Modeling and Simulation 建模与仿真, 2025, 14(5), 211-222
Published Online May 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.145387

Hans汉斯

基于特征点驱动的视觉Transformer驾驶行为分析

黄廷禾，陈庆奎*

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年4月12日；录用日期：2025年5月4日；发布日期：2025年5月13日

摘　要

针对Vision Tranformer (ViT)在局部特征捕捉和计算效率方面的局限性，文章提出了一种将目标检测技术和视觉神经网络分类模型融合到一起的方法。该方法针对驾驶行为分析特征点数量稀疏的特点，对ViT进行改进，提出了分阶段注意力计算策略，通过重构ViT编码层，将后五层的全局视觉特征序列替换为目标检测特征点序列，使其更适配基于特征点驱动的模型。并替换标准ViT的位置编码，引入方向盘相对距离－角度联合编码。此外，为了改善模型对面部捕捉的不足，在上述模型的基础上引入了多任务学习，加入了判断驾驶员面部是否直视前方这个子模型辅助判断，以更好地判断驾驶员的驾驶行为。这个模型称为特征点驱动的Vision Tranformer多任务学习模型(FViT-MTL)，在SFDDD数据集上的准确率为93.85%，比其他主流视觉神经网络分类模型提升了5.71%，比目前应用于驾驶行为分析先进的方法提升了1.28%，有效提升了分类模型的准确率，并确保驾驶行为的判断准确。

关键词

驾驶行为分析，目标检测，注意力机制，ViT

Feature Point-Driven Vision Transformer for Driving Behavior Analysis

Tinghe Huang, Qingkui Chen*

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Apr. 12th, 2025; accepted: May 4th, 2025; published: May 13th, 2025

*通讯作者。

文章引用: 黄廷禾, 陈庆奎. 基于特征点驱动的视觉 Transformer 驾驶行为分析[J]. 建模与仿真, 2025, 14(5): 211-222.
DOI: 10.12677/mos.2025.145387

Abstract

To address the limitations of Vision Transformer (ViT) in local feature capture and computational efficiency, this study proposes an integrated approach that fuses object detection technology with visual neural network classification models. Specifically, we introduce a phased attention computation strategy tailored for sparse feature point scenarios in driving behavior analysis. By reconstructing ViT's encoder layers, we replace the global visual feature sequences in the last five layers with target-detected feature point sequences, making them more suitable for feature-driven models. Additionally, we substitute standard ViT positional encoding with a combined relative position-angle encoding of the steering wheel to enhance spatial-temporal context understanding. Furthermore, to compensate for suboptimal facial feature capture, a multimodal subsystem is introduced through multitask learning to detect driver head orientation. The proposed Feature-Driven Vision Transformer Multitask Learning (FViT-MTL) achieves 93.85% classification accuracy on the SFDDD dataset, demonstrating a 5.71% improvement over conventional visual neural networks and a 1.28% gain compared to state-of-the-art methods. The results validate our approach's effectiveness in achieving both computational efficiency and precise driving behavior analysis.

Keywords

Driver Behavior Analysis, Object Detection, Attention Mechanisms, ViT

1. 引言

根据 WHO 的数据，每年约有 119 万人因道路交通事故而丧生，有 2000 万至 5000 万人因交通事故受伤。研究表明，除了道路条件和车辆技术因素外，驾驶员的行为也是引发交通事故的主要原因。开车时使用手机聊天、播放媒体、浏览网页、玩游戏、使用 GPS 或其他应用程序会导致分心驾驶从而产生交通事故[1]-[3]。

因此，如果能检测驾驶员在开车时是否有异常行为，则可以有效减少交通事故的发生率。

随着深度学习技术的快速发展，使用深度学习方法分析驾驶行为被广泛应用。先前对驾驶行为的研究主要集中于使用卷积神经网络(CNN)，Eraqi 等人[4]提出了遗传加权集成的卷积神经网络用于驾驶员分心识别，达到了 90%的准确率；Xing 等人[5]使用高斯混合模型(GMM)将原始图像分割以从背景中提取驾驶员身体，实现并评估了 AlexNet、GoogleNet 和 ResNet50 三种卷积神经网络的效果，最高达到了 81.6%的准确率。但 CNN 的方法在处理长距离依赖方面存在局限，且准确率不足。随着目标检测模型 Yolo 的兴起，越来越多的研究尝试直接使用 Yolo 检测驾驶行为[6]-[8]，然而，Yolo 是一个目标检测模型而非分类模型，不能直接给出分类结果，并存在多目标重叠场景下易产生分类歧义的情况。近年来，Vision Transformer (ViT)被引入到计算机视觉的分类任务中[9]，但 ViT 存在计算效率低和对局部特征捕捉不足的问题。Ma [10]等人的研究中将驾驶员情绪作为一种额外的模态，显著提高了检测分心行为的性能。采用双分支 ViT 架构同步提取全局场景与局部面部特征，最终准确率达 92.58%。该方法虽通过局部特征提取缓解了 ViT 的细节捕捉缺陷，但未解决其计算效率低的问题，不过也给了我们通过捕捉更多局部特征提高精度的启发。

此外，侧向拍摄数据集因其能够捕获更为丰富的驾驶员行为信息，所以可以更全面地分析驾驶行为。然而，现有侧向拍摄数据集普遍存在一个关键局限性，即未能充分考虑驾驶员面部朝向这一重要特征，驾驶员持续偏离正前方视野也是一种常见的危险驾驶行为。

为了解决上述问题，更高效地判断驾驶行为，本文提出了基于目标检测特征点驱动的 Vision Transformer 模型，结合了目标检测模型 YOLOv8 和视觉神经网络分类模型 ViT。主要贡献有：重新标注了 SFDDD 数据集，不仅为 SFDDD 标注了特征点，还新增了驾驶员是否直视正前方的标签，作为模型的一个子任务判断；将目标检测技术和 ViT 相结合，使模型能够更加集中于特征点区域，解决了 ViT 对局部特征捕捉不足的问题；提出基于特征点的编码层架构，有效减少了 ViT 的计算量；结果显示本文提出的方法在实验中准确率达到了 93.85%，优于其他方法。

2. 方法

2.1. 模型总体架构

为了更高效和准确地判断驾驶行为，本文提出了特征点驱动的 Vision Tranformer 多任务学习模型，简称 FViT-MTL。模型受到 Ma [10]等人对驾驶行为研究以及 Swin Transformer [11]窗口注意力的启发，对 ViT 在局部空间特征捕获上的局限性，本研究提出通过引入先验局部特征提取机制以增强细粒度特征表征能力，并且提出了新任务：判断驾驶员是否正视前方作为主任务(驾驶员行为分类)的辅助模态，完成多任务学习的建模。针对 ViT 计算量大的问题，受到基于 Transformer 的目标检测模型 LiTE DETR [12]的启发，提出分阶段注意力计算策略：在第一编码层采用全局特征作为键(K)和值(V)，而以特征点邻域作为查询(Q)，实现全局上下文感知；后续编码层仅使用特征点及其周围区域计算注意力。此外，引入方向盘相对距离–角度联合编码替代标准 ViT 中使用的位置编码，通过融合特征点相对于方向盘的几何位置关系增强特征点表征能力。

本文模型总体架构如图 1 所示。该模型分为两个部分：主任务和子任务的分支，其中主任务又包括对局部特征的提取和对全局特征的提取。

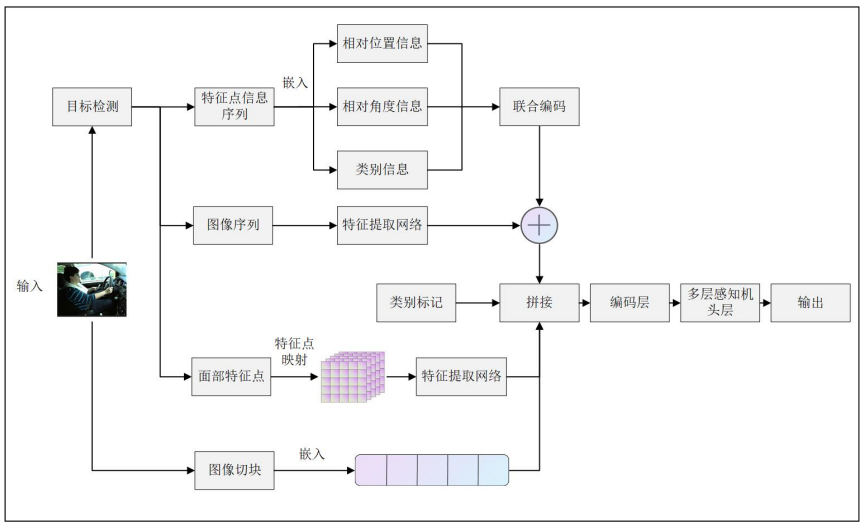


Figure 1. FViT-MTL model architecture
图 1. FViT-MTL 模型架构

主任务使用 YOLOv8 进行目标检测，提取包括驾驶员的人体、面部，方向盘，水杯，手机的特征点。提取到的特征点及其周围区域随后被构建为图像序列，输入至特征提取网络(主干网络)，并加上方向盘相

对距离－角度联合编码。方向盘相对距离－角度联合编码是各个特征点相对于方向盘的位置、角度以及自己本身的类别信息经过嵌入表处理后的编码。将方向盘作为基准点的原因是方向盘的位置相对于驾驶员的身体部位更加稳定，不会随着驾驶员的动作而变化。在自动驾驶领域中，就常以当前驾驶员驾驶的车辆作为基准点，使用车载传感器进行数据采集和处理，计算、识别其他物体的相对距离、相对速度和类别以完成辅助驾驶，是智能交通系统的核心组成部分[13]。方向盘相对距离角度－联合编码可以很好地捕捉到几个特征点之间的相对位置关系，能够有效地反映出驾驶员的动作特征以及动作间的关系。

对全局特征提取沿用标准 ViT 流程，将输入图像分割为 16 × 16 块(patch size = 14 × 14)，经线性投影获得 $H = 196$ 的序列。

针对视线方向判别子任务，提出稀疏热图编码方法：将面部 5 个特征点映射至 80 × 80 网格，通过特征增强生成五通道热图(对应眼、鼻、嘴部区域)的稀疏色块图，这种方法的特征图特征信息更明显、特征图更稀疏、收敛较快。之后这张稀疏色块图会以和主任务相同的特征提取网络提取特征。

提取完局部特征、全局特征及子任务特征之后，所有特征向量被拼接，并输入至基于分阶段注意力计算策略的编码层。该编码层能够同时关注局部和全局特征，有助于模型在学习关键特征的同时保留全局信息，从而提升模型的泛化能力。此外，通过分阶段计算，编码层有效减少了模型的计算负担，使得实时驾驶行为检测成为可能。本文的嵌入向量维度统一设置为 768。

2.2. 图像序列和特征点信息序列

通过目标检测技术，我们可以获得特征点及其周围区域的图像序列和特征点信息序列输入图像(224 × 224 × 3)，采用非重叠分块策略切分为 196 个 16 × 16 × 3 的子块，形成离散化特征定位的几何基础，这些块定义为 p_{ij}(i 和 j 表示行列索引)。不同于传统 ViT 的卷积切块，因为本文最终要获取的是特征点及其邻近的图像区域，切块是为了获取离散特征点的位置信息以便于位置向量的嵌入，并通过块来得到特征点的邻近图像区域。特征点范围∈[“驾驶员面部”，“驾驶员左手部分”，“驾驶员右手部分”，“方向盘”，“手机”，“水杯”]。以检测到的特征点为中心，提取 5 × 5 邻域的 80 × 80 × 3 图像块(不足区域补 0)，称为 P^k。生成特征点及其周围区域的图像序列。该序列长度 n 为特征点数量，每个图像块经特征提取网络处理后，输出形状为 n × 1 × 768 的特征向量序列，称为 S_1：

$$P^k = P_{ij}^k = \begin{bmatrix} p_{i-2\,j-2}^k & \cdots & p_{i+2\,j-2}^k \\ \vdots & p_{ij}^k & \vdots \\ p_{i-2\,j+2}^k & \cdots & p_{i+2\,j+2}^k \end{bmatrix} \quad k = 1, 2 \cdots n \tag{1}$$

$$S_1 = \left[backbone\left(P^1\right), backbone\left(P^2\right), \cdots, backbone\left(P^n\right) \right] \tag{2}$$

Figure 2. Image sequence
图 2. 图像序列

图像序列示意图如图 2 所示。

特征点信息序列是组成方向盘相对距离－角度联合编码的部分。包含了特征点的位置信息、角度信息和类别信息。方向盘相对距离－角度联合编码代替了位置编码，帮助模型理解输入序列中元素的位置、角度和类别的综合关系。方向盘相对距离－角度联合编码通过嵌入表的方式将离散数据转为连续向量表示：

距离编码：定义特征点到方向盘的相对位置为特征点位置至方向盘位置需要穿越的块数，嵌入表大小为 29 × 768 (因为图像被切分为 14 × 14 的块，即最大距离为 14 + 14)。

角度编码：定义特征点到方向盘的连线相对于横轴的角度，嵌入表大小为 361 × 768。

类别编码：将 6 类特征点映射至向量，嵌入表大小为 7 × 768。

为确保批次间特征索引的规范性与嵌入过程的数值稳定性，对离散特征进行偏移编码处理：将原始索引从 0-based 调整为 1-based (即输入维度增加 1)，从而避免零填充区域对嵌入矩阵的干扰。所以位置、角度及类别编码的取值范围分别为[1, 29]、[1, 361]和[1, 7]，对应的嵌入表维度相应扩展为$(m + 1) \times 768$ (其中 m 为特征取值范围)。

嵌入后使用堆叠－一维卷积的方法将信息融合，合成方向盘相对距离－角度联合编码序列，称为 S_2。与标准 ViT 相同，编码直接与特征点的特征向量进行加法操作，将计算完成的序列称为 S_{fea}。

$$S_{fea} = S1 + S2 \tag{3}$$

最后，拼接可学习的类别标记。类别标记在整个网络中起到了分类的作用，在模型的最后，类别标记所对应的输出特征会被用来进行分类任务。使用类别标记的原因是类别标记是最后加入特征序列中的，它不会受其他特征的干扰，可以完成对序列中其他特征的信息聚合，并最后用于分类任务。

2.3. 稀疏热图编码方法

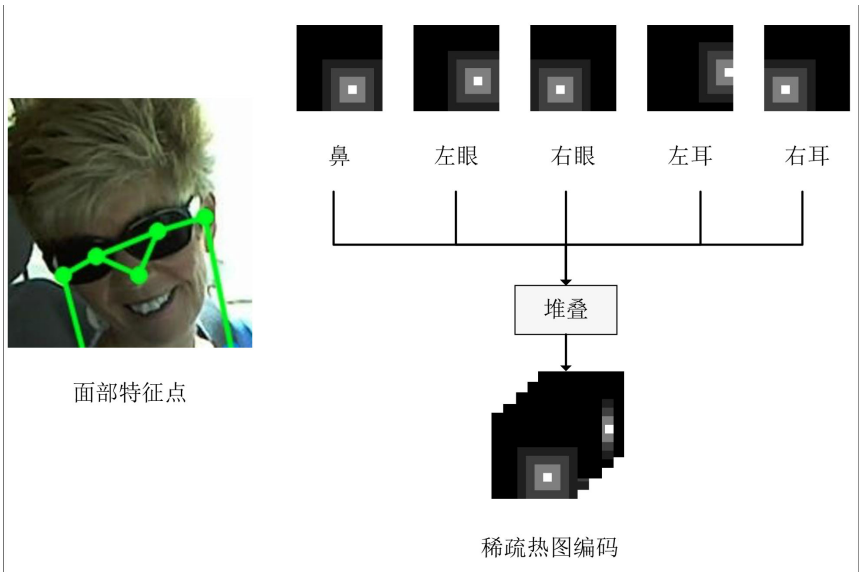


Figure 3. Input form of the facial orientation deviation judgment model
图 3. 面部朝向偏离判断模型输入形式

在判断面部是否直视正前方的子任务中，首先改进了输入形式。若直接输入面部图像，对五官之间的相对位置关系不明确，特征点与周围信息的区分度不高，因此我们提出一种基于五官特征点的特征增

强策略：稀疏热图编码方法。本文提出的方法受到了 Gupta [14]等人通过五官可能位置的热力图预测人脸朝向的回归值方法的一定启发，在此基础上提出了更适合应用于特征点驱动的输入方式。具体做法是将预测的面部特征点(即五官)位置投影到一张五通道的图中，每个特征点作为一个颜色为(255)的块投影到一张 80 × 80 像素、颜色全为(0)的黑底灰度图中。黑色底片划分为 10 × 10 的网格，每个网格块的大小为 8 × 8 像素，根据面部特征点在人脸图像中的相对位置，将黑底图片中对应网格块的颜色改为(255)，即白色。为了增强特征图的特征信息，我们进一步处理基准块的周围区域。将基准块外围一圈的块颜色填充(128)，即灰色，这样称为一次特征增强。每次特征增强的颜色值都会减半。

通过这种方式，保留了特征点的位置信息和特征点之间的相对位置关系，并增强了特征点信息的视觉表示，使模型能够更好地学习五官的相对位置和面部朝向之间的关联。本文提出的模型特征增强次数为三次，具体的输入形式如图 3 所示。

将五通道热图使用特征提取网络提取特征，拼接进特征序列。此时特征序列的长度为$(n + 2) \times 768$，其中 n 为特征点的数量。

2.4. 编码层

编码层中的多头自注意力计算采用本文提出的分阶段注意力计算策略计算。编码层由多头自注意力机制和一个多层感知机组成，并在多头自注意力机制和多层感知机之间进行层归一化；之后应用残差链接，以缓解梯度消失问题。为了更精确地提取特征以及处理序列之间的关系，编码层需要堆叠 6 次，编码层的结构如图 4 所示。分阶段注意力计算采用层级化的特征交互策略：在第一编码层采用全局特征作为键(K)和值(V)，而以特征点邻域作为查询(Q)，实现全局上下文感知；后续编码层仅使用特征点及其周围区域计算注意力。全局特征采用标准分块策略进行特征提取：将输入图像切分为 16 × 16 的非重叠块，经嵌入层处理后生成形状为 196 × 768 的全局特征向量。分阶段注意力计算策略在注意力专注于特征点及其周围区域的同时，也融合了全局特征，实现了更多特征信息的融合，增加了模型的泛化能力。分阶段注意力计算策略通过聚焦于特征点，不仅解决了 ViT 无法准确捕捉局部特征的缺点，也通过减少序列的长度，减少了计算注意力时的计算量。

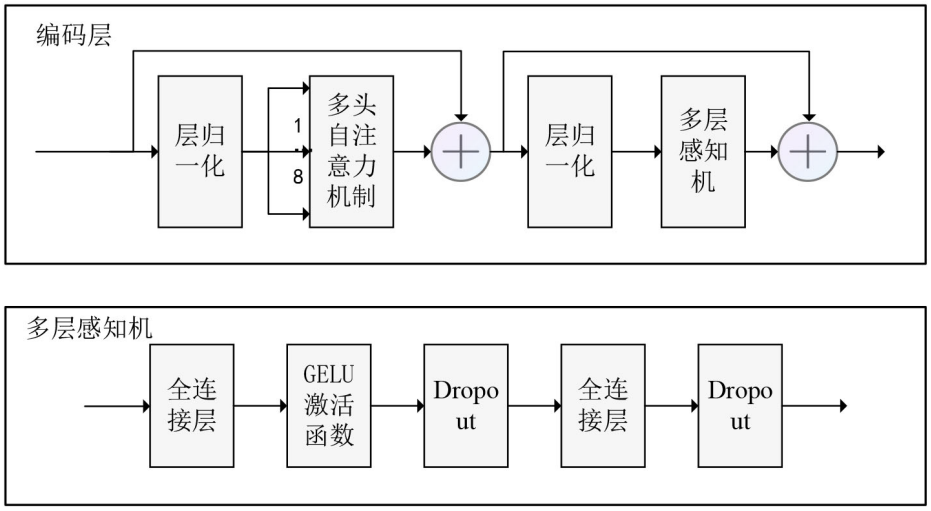


Figure 4. Encoder layer architecture
图 4. 编码层架构

其中，分阶段注意力计算策略如图 5 所示。

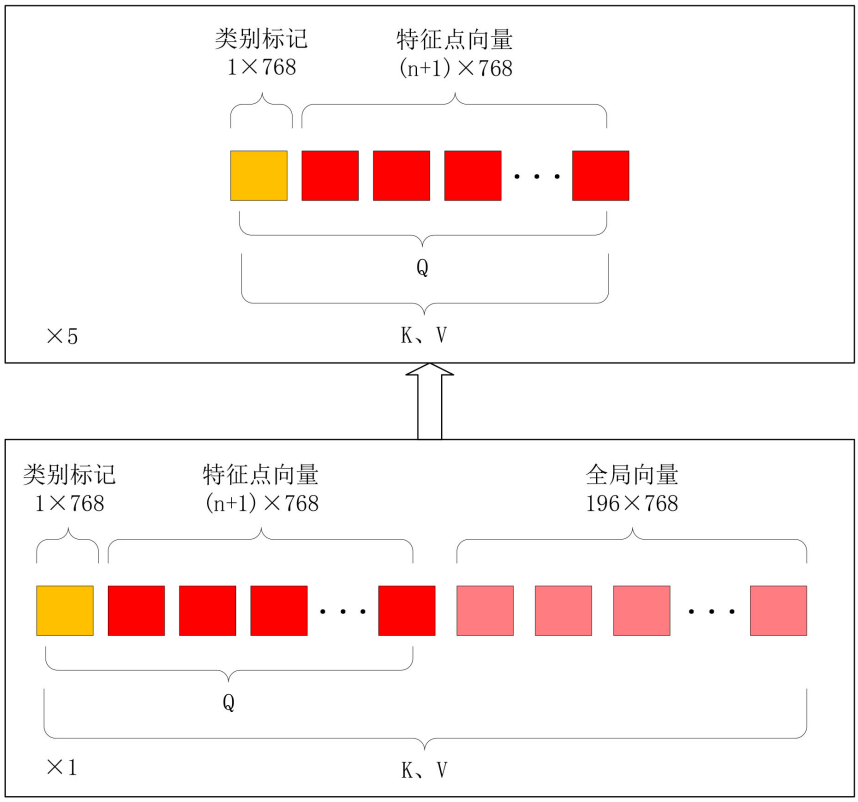


Figure 5. Phased attention computation strategy
图 5. 分阶段注意力计算策略

3. 实验

3.1. 数据集

StateFarm 的分心驾驶员数据集(SFDDD)是主流的驾驶行为检测数据集，它将驾驶行为分为了 10 类，包括安全驾驶、右手发短信、右手打电话、左手发短信、左手打电话、操作收音机、喝水、伸手到后面、化妆、和后座乘客对话。需要注意的是，驾驶员面部异常的范围不仅限于 SFDDD 数据集中定义的行为，还包括许多其他异常行为(如打哈欠，面部转动等)，但这些在当前数据集中未涵盖。我们在数据集的基础上标注了驾驶员面部朝向是否直视前方，并将其作为子任务之一进行建模。

SFDDD 数据集已被广泛应用于驾驶行为分析的研究中。数据集由 26 名驾驶员的 22,424 张有类别的图片组成，分辨率为 640 × 480。现有研究采用两种数据划分方式：一种按照图片划分训练集与测试集，直接将图片的 80%作为训练集，剩余 20%作为测试集，这种划分方式在前人的研究中被广泛使用[15] [16]，但其局限性是训练集和测试集的强相关性无法准确量化模型的性能。另一种方式是按驾驶员划分训练集与测试集，确保同一人不会同时出现在训练集和测试集。本文的训练方法是将 80%的驾驶员(21 人)作为训练集，剩余 20%的驾驶员(5 人)作为测试集，有效缓解数据泄露问题，提升模型泛化能力评估可信度。

3.2. 实验环境

本文使用的操作系统是 Linux 5.15.0-107，Ubuntu 20.04.2，Python 版本是 3.12.2，PyTorch-CUDA 版本是 11.8。在硬件设备上，GPU 是 NVIDIA GeForce RTX 3080 Ti，显存 12 GB，CPU 是 Intel (R) Xeon

(R) CPU E3-1231 v3，内存为 32 GB。

训练中采用的优化器是 AdamW 优化器，权重衰减为 0.1。AdamW 的优势是在 Adam 的基础上加入权重衰减。Adam 优化器实现权重衰减的方法是在梯度上添加正则项，再进行反向传播；AadmW 则是直接在参数更新时加入权重衰减，可以更好地实现正则化的效果，防止过拟合。训练驾驶行为分类模型使用余弦退火调度器，在前五个轮次中学习率从 1e−6 上升到最大学习率 0.0002，在之后 95 个轮次中平滑地降低到接近零的值。训练轮数为 100，批次大小为 16。

3.3. 对比实验

为了准确地衡量 FVIT-MTL 模型的性能，我们与其他模型进行了对比实验。对比的模型包括 Resnet50、DenseNet121、VGG16、ViT_L_16、Swin Transformer 以及 VIT-DD。对比实验的评价指标均为准确率与 F1 指数，其中准确率可以直观地在分类任务中衡量模型的性能，F1 指数是精度和召回率的调和平均数，可以更全面地表示模型的性能。对比实验中 FVIT-MTL 的特征提取网络使用的是 ViT。

对比实验结果如表 1 所示。

Table 1. Comparative experimental results
表 1. 对比实验结果

模型	Accuracy	F1
ResNet50	83.68	85.32
DenseNet121	88.14	89.78
VGG16	79.19	79.44
ViT_b_16	50.97	51.10
Swin Transformer	56.68	56.81
VIT-DD	92.57	/
FViT-MTL	93.85	93.80

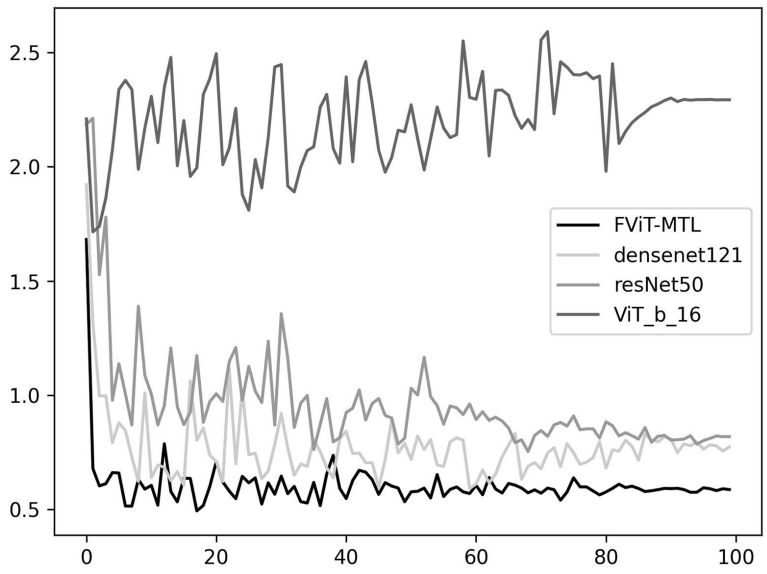


Figure 6. Comparison of loss changes
图 6. 损失变化对比

实验结果表明，我们提出的模型不论是在准确率还是在 F1 指数上都领先于其他主流视觉神经网络分类模型。在驾驶行为分类模型中，模型比主流模型中准确率最高的 DenseNet121 提升了 5.71%的准确率，比标准的 ViT_b_16 和 Swin Transformer 分别提升了 43.9%和 37.17%的准确率，比目前先进的应用于 SFDDD 数据集的方法 VIT-DD 准确率高出 1.4%。这表明了通过聚焦于特征点区域，我们提出的模型能更有效地捕捉图像中的重要特征，减少无关背景的干扰，从而提高 ViT 的模型性能。

ViT_b_16 和 Swin Transformer 均无法在 100 个 epoch 内收敛，突出了这两个模型计算效率低、无法聚焦于局部特征的缺陷。FViT-MTL 通过结合 YOLOv8 检测的特征点，相比标准 ViT 能够更高效地捕获信息，避免了 ViT 在计算注意力时可能出现的缺乏对关键区域有效捕获的问题，从而实现训练过程的快速收敛，并且在实际效果上表现优异。FViT-MTL、ResNet50、DenseNet121 和 ViT_b_32 四个模型训练过程中，测试集的损失变化如图 6 所示，图中横坐标代表训练轮数 Epoch，纵坐标代表损失值 Loss。

从损失曲线图中可以观察到，FViT-MTL 相较于其他模型具有更快的收敛速度，表现出更高的训练效率。同时，其损失值的波动幅度更小，反映出模型在训练过程中的稳定性更高。这表明 FViT-MTL 在优化过程中能够更迅速地找到全局最优解，同时有效避免了过大的梯度波动所导致的不稳定现象。

3.4. 消融实验

FVIT-MTL 在处理图像序列和映射的面部特征点热图时使用了特征提取网络(主干网络)，通过消融实验，可以确认特征提取网络对模型的综合影响。主干网络的选择评价指标包括准确率与 F1 指数。主干网络选择对比如表 2 所示。

Table 2. Comparison of backbone network selection
表 2. 主干网络选择对比

模型	Accuracy	F1
ResNet18	91.43	91.37
ResNet50	92.56	92.41
ConvNext	93.75	93.71
ViT	93.85	93.80

从对比结果可以看出，ViT 是最适合 FViT-MTL 特征提取网络的架构，准确率在四个网络中最高。

为了验证通过将方向盘距离－角度编码替换标准 ViT 位置编码的有效性，将 FViT-MTL 中的方向盘距离－角度编码替换为标准 ViT 的位置编码进行消融实验，消融实验结果如表 3 所示。

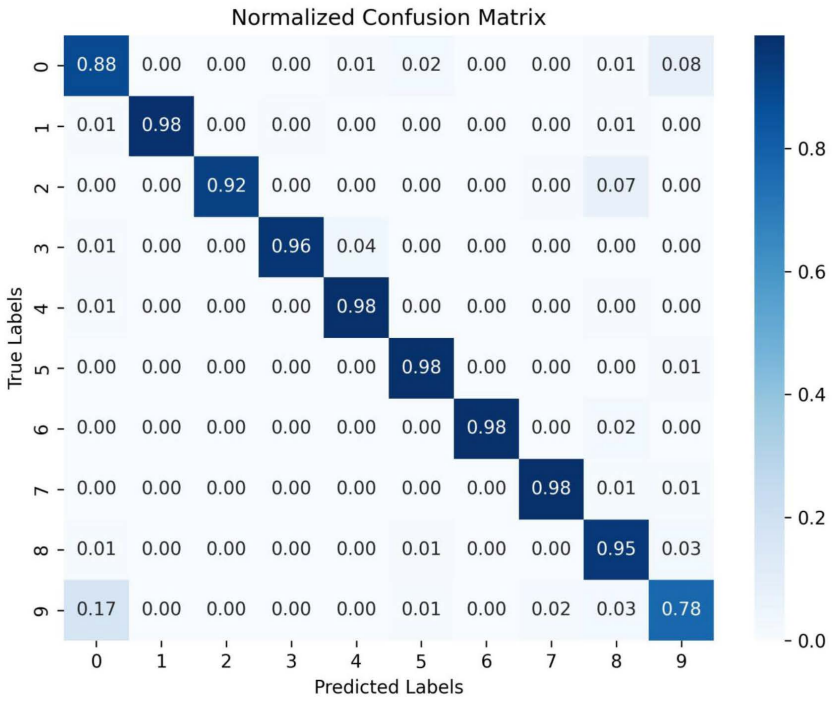
Table 3. Ablation experiment on steering wheel distance-angle encoding
表 3. 方向盘距离－角度编码消融实验

编码	Accuracy	F1
ViT 位置编码	78.66	78.14
方向盘距离－角度编码	93.85	93.80

将方向盘距离－角度编码替换回标准 ViT 编码后，模型对于各个特征点之间的相对位置关系的学习不够充分，准确率降低了 15.66%。这验证了我们提出的方向盘距离－角度编码方法的有效性以及对于 FViT-MTL 模型的必要性。该方法整合了各特征点相对于方向盘的距离信息、角度信息以及自身类别信息，从而有效地增强了模型对驾驶行为分类中各特征点相对位置关系的学习能力。

本研究构建了消融实验框架以验证多任务学习的有效性：将单任务模型称为 FViT，FViT 聚焦于特征点间的相对位置关系建模，对于第十类(驾驶员与乘客对话)这类需要专注于驾驶员面部姿态类别分类的效果较差。FViT-MTL 与 FViT 分类结果的混淆矩阵如图 7 所示。

FViT-MTL混淆矩阵



FViT混淆矩阵

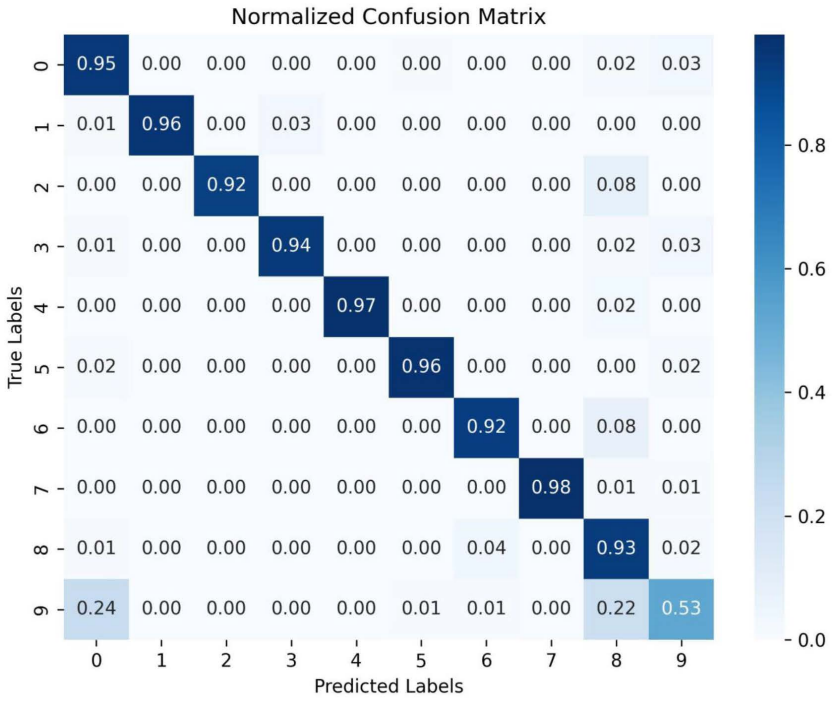


Figure 7. Confusion matrix of FViT-MTL and FViT classification results
图 7. FViT-MTL 与 FViT 分类结果的混淆矩阵

实验结果表明，多任务学习相比单任务学习提升了 3.5%的准确率，尤其在第十类(驾驶员与乘客对话)中提升明显，准确率提升了 26%。

为验证在子任务处理中将面部五官信息转换为稀疏热图编码方法的有效性，本文设计并构建了消融实验。实验对比了直接输入原始图像与将原始图像中的五官信息转换为稀疏热图编码两种方法的性能。消融实验结果如表 4 所示。

Table 4. Ablation experiment on sparse heatmap encoding
表 4. 稀疏热图编码消融实验

子任务处理方法	Accuracy (Main)	F1 (Main)	Accuracy (Sub)	F1 (Sub)
原图输入	90.63	90.39	90.29	87.68
稀疏热图编码	93.85	93.80	98.96	97.38

从消融实验结果可以看出，当直接使用驾驶员面部原始图像作为输入时，针对驾驶员面部是否正对前方这一子任务的准确率仅为 90.29%，且对主任务的准确率产生了负面影响。与采用本文提出的稀疏热图编码方法相比，直接输入原图使得子任务和主任务的判断准确率分别下降了 8.67%和 3.22%。该结果表明，所提出的模型能够通过捕捉面部五官之间的相对位置关系，更准确地判断驾驶员面部朝向是否正对前方，从而实现对驾驶行为的更全面检测。

4. 结语

本文提出的基于目标检测特征点驱动的 Vision Transformer 驾驶行为分析方法，将目标检测技术融入到视觉神经网络分类模型中，建立 FViT-MTL 模型，并采取方向盘相对距离－角度编码、稀疏热图编码方法以及分阶段注意力策略，有效地捕捉了图像特征点的特征信息及其之间的相对关系，使分类精度得到提高。实验证明，本文提出的方法在 SFDDD 数据集上取得了 93.85%的分类准确率，优于其他模型，可以完成对驾驶行为的有效分析和判别。

在未来的研究中，若能减少目标检测模型的参数量，或提升目标检测模型的精度，可以让我们的方法在实验中取得更加显著的成果。

在实际驾车场景中，若能将这项技术率先应用到公共交通(如公交车、出租车)中，可以有效判断驾驶员是否有危险驾驶行为，从而降低交通事故的发生概率。

基金项目

国家自然科学基金项目(61572325)；上海重点科技攻关项目(19DZ1208903)。

参考文献

[1] Khandakar, A., Chowdhury, M.E.H., Ahmed, R., Dhib, A., Mohammed, M., Al-Emadi, N.A.M.A., *et al*. (2019) Portable System for Monitoring and Controlling Driver Behavior and the Use of a Mobile Phone While Driving. *Sensors*, **19**, Article 1563. https://doi.org/10.3390/s19071563

[2] Coxon, K. and Keay, L. (2015) Behind the Wheel: Community Consultation Informs Adaptation of Safe-Transport Program for Older Drivers. *BMC Research Notes*, **8**, Article No. 764. https://doi.org/10.1186/s13104-015-1745-0

[3] Lechner, G., Fellmann, M., Festl, A., Kaiser, C., Kalayci, T.E., Spitzer, M., *et al*. (2019) A Lightweight Framework for Multi-Device Integration and Multi-Sensor Fusion to Explore Driver Distraction. In: Giorgini, P. and Weber, B., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 80-95. https://doi.org/10.1007/978-3-030-21290-2_6

[4] Eraqi, H.M., Abouelnaga, Y., Saad, M.H. and Moustafa, M.N. (2019) Driver Distraction Identification with an Ensemble of Convolutional Neural Networks. *Journal of Advanced Transportation*, **2019**, Article 4125865.

https://doi.org/10.1155/2019/4125865

[5] Xing, Y., Lv, C., Wang, H., Cao, D., Velenis, E. and Wang, F. (2019) Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach. *IEEE Transactions on Vehicular Technology*, **68**, 5379-5390. https://doi.org/10.1109/tvt.2019.2908425

[6] Du, Y., Liu, X., Yi, Y. and Wei, K. (2023) Optimizing Road Safety: Advancements in Lightweight YOLOv8 Models and GhostC2f Design for Real-Time Distracted Driving Detection. *Sensors*, **23**, Article 8844. https://doi.org/10.3390/s23218844

[7] 周宏威, 王文博, 王伟光, 等. 基于改进 YOLOv7 算法的驾驶分心行为检测模型[J/OL]. 自动化技术与应用, 1-10. http://kns.cnki.net/kcms/detail/23.1474.TP.20241230.0914.028.html, 2025-02-19.

[8] 王宝峰. 基于 Faster-YOLO 模型的驾驶员行为检测方法研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2024.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2021) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *The Ninth International Conference on Learning Representations*, *ICLR* 2021, Virtual Event, 3-7 May 2021, 1-21.

[10] Ma, Y., Du, R., Abdelraouf, A., Han, K., Gupta, R. and Wang, Z. (2024) Driver Digital Twin for Online Recognition of Distracted Driving Behaviors. *IEEE Transactions on Intelligent Vehicles*, **9**, 3168-3180. https://doi.org/10.1109/tiv.2024.3353253

[11] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

[12] Li, F., Zeng, A., Liu, S., Zhang, H., Li, H., Zhang, L., *et al*. (2023) Lite DETR: An Interleaved Multi-Scale Encoder for Efficient DETR. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 18558-18567. https://doi.org/10.1109/cvpr52729.2023.01780

[13] Li, K., Xiong, H., Liu, J., Xu, Q. and Wang, J. (2022) Real-Time Monocular Joint Perception Network for Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems*, **23**, 15864-15877. https://doi.org/10.1109/tits.2022.3146087

[14] Gupta, A., Thakkar, K., Gandhi, V. and Narayanan, P.J. (2019) Nose, Eyes and Ears: Head Pose Estimation by Locating Facial Keypoints. *ICASSP* 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 1977-1981. https://doi.org/10.1109/icassp.2019.8683503

[15] Khan, T., Choi, G. and Lee, S. (2023) EFFNet-CA: An Efficient Driver Distraction Detection Based on Multiscale Features Extractions and Channel Attention Mechanism. *Sensors*, **23**, Article 3835. https://doi.org/10.3390/s23083835

[16] Hossain, M.U., Rahman, M.A., Islam, M.M., Akhter, A., Uddin, M.A. and Paul, B.K. (2022) Automatic Driver Distraction Detection Using Deep Convolutional Neural Networks. *Intelligent Systems with Applications*, **14**, Article 200075. https://doi.org/10.1016/j.iswa.2022.200075