计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2024,33(11):1−14 [doi: 10.15888/j.cnki.csa.009677] [CSTR: 32024.14.csa.009677]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多模态数据融合的飞行员注视区域分类①

段高乐[1], 王长元[1], 吴恭朴[2], 王红艳[1]

[1](西安工业大学 计算机科学与工程学院, 西安 710021)
[2](西安工业大学 光电工程学院, 西安 710021)
通信作者: 王长元, E-mail: cyw901@163.com

摘　要: 为了解决图像采集过程中眼图消失和头部姿态估计不准确的问题, 利用基于非接触式的眼部信息获取方法采集人脸图像, 从单个图像帧中确定飞行员当前的注视方向. 同时, 针对现有网络忽略头部运动对视线造成遮挡所导致的分类效果不佳问题, 结合人脸图像与头部姿态特征, 通过改进的 MobileViT 模型提出一种用于飞行员注视区域分类的多模态数据融合网络. 首先提出了多模态数据融合模块解决特征拼接过程中尺寸不平衡导致的过拟合问题, 其次提出一种基于并行分支 SE 机制的逆残差块, 充分利用网络浅层的空间和通道特征信息, 并结合 Transformer 的全局注意力机制捕捉多尺度特征. 最后, 重新设计了 Mobile Block 结构, 使用深度可分离卷积降低模型复杂度. 利用自制数据集 FlyGaze 对新模型和主流基线模型进行对比, 实验结果表明, PilotT 模型对注视区域 0、3、4、5 的分类准确率均在 92% 以上, 且对人脸发生偏转的情况具有较强适应力. 研究结果对提升飞行训练质量以及飞行员意图识别和疲劳评估具有实际应用价值.

关键词: 注视区域分类; 并行分支 SE 机制; MobileViT; 多模态数据融合

引用格式: 段高乐,王长元,吴恭朴,王红艳.基于多模态数据融合的飞行员注视区域分类.计算机系统应用,2024,33(11):1–14. http://www.c-s-a.org.cn/1003-3254/9677.html

Pilot’s Gaze Zone Classification Based on Multi-modal Data Fusion

DUAN Gao-Le[1], WANG Chang-Yuan[1], WU Gong-Pu[2], WANG Hong-Yan[1]

[1](School of Computer Science and Engineering, Xi’an Technology University, Xi’an 710021, China)
[2](School of Optoelectronic Engineering, Xi’an Technology University, Xi’an 710021, China)

Abstract: To avoid eye image disappearance and inaccurate head pose estimation during image capture, a non-contact method for acquiring eye information is employed to collect facial images, determining the pilot’s current gaze direction from a single image frame. Concurrently, considering the poor classification of current networks due to the neglect of visual obstruction caused by head movements, with a combination of facial images and head poses, a multimodal data fusion network for the pilot’s gaze region classification is proposed using an improved MobileVit model. Firstly, a multi-modal data fusion module is introduced to address the problem of overfitting resulting from size imbalances during feature concatenation. Additionally, an inverse residual block based on a parallel branch SE mechanism is proposed to fully leverage spatial and channel feature information in the shallow layers of the network. Moreover, multi-scale features are captured by integrating the global attention mechanism from the Transformer. Finally, the Mobile Block structure is redesigned and the depthwise separable convolution is utilized to reduce model complexity. Experimental comparisons with mainstream baseline models are conducted using a self-made dataset FlyGaze. The results demonstrate that the PilotT model achieves classification accuracies exceeding 92% for gaze regions 0, 3, 4, and 5, with robust adaptability to facial deflection. These findings hold practical significance for enhancing flight training quality and facilitating pilot

① 基金项目: 国家自然科学基金 (52072293)
收稿时间: 2024-04-21; 修改时间: 2024-05-20; 采用时间: 2024-06-04; csa 在线出版时间: 2024-09-27
CNKI 网络首发时间: 2024-09-29

intention recognition and fatigue assessment.

Key words: gaze region classification; parallel branch SE mechanism; MobileViT; multi-modal data fusion

随着计算机视觉、深度学习以及摄像头传感器硬件的快速发展, 基于特征学习的注视区域分类方法[1]取得了显著进展, 这类方法利用深度学习网络从大规模数据中学习特征表示, 为实现高度智能化的注视追踪系统提供了有力支持[2,3], 在人机交互和辅助驾驶等领域具有广泛的应用前景, 特别是在航空领域, 对注视区域进行分类有助于评判飞行训练中飞行学员的操作是否符合任务标准. 通过分析注视分布状况, 可以发现错误的操作并及时纠正, 从而大大提升飞行训练的质量. 除此之外, 通过实时监测和分析飞行员的视线分布, 可以监测飞行员的疲劳状态以提高实际飞行操作的安全性和效率. 随着相关技术的进一步成熟, 注视区域分类方法有望在更多领域中得到应用, 推动智能化人机交互系统的全面发展.

然而达成这一目标所面临的挑战是从图像和视频中确定人类视觉系统优先关注的场景区域, 这些显著区域通常包含关键的场景或目标信息, 对于理解人的注意焦点至关重要. 但复杂的头部运动、视角方向都会导致注视方向发生变化. 现有方法大多通过在图像中加入网格或直接利用图像中头部朝向信息进行隐式的头部姿态估计, 并未进行准确的头部姿态解算. 结合头部姿态数据有助于提高注视区域分类的准确性, 在融合人脸图像和头部姿态数据的背景下, 基于卷积神经网络 (convolutional neural network, CNN) 的模型忽略了 2 种特征之间的相关性, 网络浅层无法捕获全局信息, 导致分类效果不佳[4]. 虽然基于 Transformer 的模型仅依赖于级联注意力机制或卷积运算, 可以捕获特征空间中的长距离依赖关系, 但其缺乏对局部特征信息的捕获, 导致浅层特征信息丢失.

针对以上问题, 利用基于非接触式系统的眼部信息获取方法[5], 通过改进 MobileViT 的多模态数据融合飞行员注视区域分类网络 (pilot Transformer gaze region classification network, PilotT) 对飞行员在飞行过程中的注视区域进行准确分类. 具体而言, 设计了一种基于并行分支 SE 注意力机制[6]逆残差块[7], 在 Mobile Block 中使用深度可分离卷积 (depthwise separable convolution, DW)[8]替代普通卷积. 同时提出一种多模态数据融合 (multi-modal data fusion module, MDFM) 模块, 结合编码器各阶段的人脸图像特征, 通过卷积层和全连接层将头部姿态特征与人脸图像特征进行融合. 经实验证明, PilotT 在图像信息处理方面表现出色, 特别是在应对头部姿态变化及视线方向变化时具备独特的优势. 主要贡献可以总结为以下 4 个方面.

(1) 提出了 MDFM 模块, 充分利用图像特征和头部姿态特征, 补偿头部运动导致的视线遮挡误差, 并解决了特征拼接过程中特征尺寸不平衡导致的过拟合问题.

(2) 设计了一种基于并行分支 SE 机制的逆残差块, 双分支分别关注通道特征间的关系以及特征图的空间分布, 利用浅层特征提升网络分类决策能力, 并在 Mobile Block 结构中使用 DW 卷积替换普通卷积, 从而降低模型复杂度.

(3) 构建了基于模拟飞行环境下飞行员视线区域数据集 FlyGaze. 目前针对飞行员视线区域分类研究的公开数据集很少, FlyGaze 数据集将飞行员的注意力划分为 6 个主要区域, 为进一步的意图识别和疲劳评估研究提供了真实场景下的人脸图像和头部姿态数据.

(4) 对比分析了 MobileNetV2/V3[9]、Swin Transformer (SwinT)[10]、Vision Transformer (ViT)[11]、MobileViT[12]、ResNet50[13]、EfficientNet[14]、ConvNeXt[15]等主流基线模型在有无头部姿态数据 2 种场景下的分类效果, 同时通过消融实验证明了模型改进部分的有效性.

1 相关工作

随着人工智能技术的发展, 高度智能化的人机协同驾驶逐渐成为现实. 人机协同驾驶需要准确识别并预测驾驶员的行为, 以更好分配人机间的驾驶权, 从而构建高效的人机交互系统. 从人与传感器之间的关系来看, 目前用于注视区域分类的方法通常分为接触式方法和非接触式方法. 接触式方法通常利用穿戴式眼动仪等设备对瞳孔、角膜中的生理变化进行跟踪并记录头部运动共同推断注视方向. 但在实际使用过程中, 该方法容易使被试感到疲惫, 从而可能造成潜在的风

险. 非接触式系统主要通过布置在被试周围的摄像头来捕获面部图像, 继而从图像中确定注视方向和注意力状态, 该方法不会对被试造成任何干扰, 成本低且便于部署, 应用前景更加广泛.

目前基于非接触式传感器的注视区域分类算法根据特征的差异主要分为 3 类: 基于人脸特征算法[16], 基于头部姿态算法[17]以及人脸特征与头部姿态相结合的算法[18]. Cheng 等[19]根据双眼在光照和头部运动下表现出的不对称性, 利用非对称回归评价策略从左右眼中选择特征信息更多一方, 增加该眼睛在网络中的权重以提高注视区域分类效果, 但该方法去除了人脸图像中的头部姿态, 仅将人脸和双眼图像作为输入, 在低光环境和头部运动幅度较小的情况下分类效果较差. 文献[20,21]表明, 除眼睛以外的其他面部区域同样包含有价值的注视信息, 针对面部区域中不必要的区域信息, Liu 等[22]通过人脸检测器提取人脸图像, 利用多通道和空间注意力神经网络增强人脸图像中的重要特征, 并抑制不必要的特征, 但在图像中眼睛部分产生缺失和模糊的情况下效果较差. Balim 等[23]跳过了人脸和眼部的预处理部分, 组合未经裁剪的上半身图像中的 2D 注视原点位置和稀疏深度图来生成 3D 注视原点, 并通过多层感知机确定 3D 注视方向, 但该方法使用的图像中有效面部尺寸较小, 未捕捉充分的面部特征信息, 导致精度不足, 鲁棒性较差. 针对人与相机之间的不同距离、头部姿态的差异与眼镜或头发遮挡的问题, Dai 等[24]将融合的双目特征和根据人脸网格定位的头部位置信息作为模型输入, 在 ResNet 的基础上引入局部与全局双目空间注意力机制以获取注视方向, 结果表明双目特征融合性能更好, 但该方法只能处理正面人脸图像, 无法通过侧面人脸图像获取注视方向.

Cazzato 等[25]总结了近年来计算机视觉和深度学习在注视估计领域的显著进展, 并强调了单一度量标准的局限性, 仅靠人脸或头部特征难以准确估计由头眼运动共同导致的注视方向变化, 由此提出将人脸特征和头部姿态特征相结合的方法. 戴忠东等[26]针对单目眼睛图像易受头部运动而失真的问题, 利用 3D 人脸模型和相机内参建立头部姿态坐标系, 使用归一化坐标系校正相机坐标系以复原人眼图像, 并通过黄金分割搜索策略进一步优化分类效果, 但使用高分辨率图像时模型训练效果较差, 且在低光照环境下效果欠佳. 闫秋女等[27]使用 POSIT 算法解算得到驾驶员头部姿态, 并采用基于 3D 人眼模型的方法由面部 2D 关键点估计注视方向, 进而融合头部姿态及注视方向特征进行分类, 但分类准确率过于依赖人脸关键点的检测精度, 在眼睛闭合情况下精度较差.

将深度学习应用在注视区域分类领域是当下的研究热点. Ghosh 等[28]以 Inception-V1 为骨干网络, 并加入光照鲁棒层对 RGB 人脸图像进行处理, 提取人脸注视特征, 但未评估头部姿态信息对网络的有效性, 且网络泛化性较差. Ali 等[29]提出一种多流 3D 注视分类网络, 不需要单独提取头部姿态数据, 输入双眼图像和人脸位置, 利用基线 CNN 进行多通道特征提取, 输出注视分类结果, 该方法可以有效提高分类准确率, 但模型参数量庞大, 模型结构过于复杂. 杨易蓉[30]通过 Kronecker 内积和空间注意力机制融合头眼特征, 实现将驾驶员的待评估视线映射在交通场景的大致区域上, 该方法未建立视线与场景内目标端到端的映射, 而是使用标定转换方法采集数据, 且受摄像头安装位置限制, 注视部分区域时的眼部可能会被遮挡. 张名芳等[31]利用多任务级联 CNN 进行人脸对齐, 构建循环生成对抗网络移除眼镜遮挡, 并将 RGB 图像输入基于 MobileNetV2 的循环神经网络, 输出注视区域分类结果, 但 RGB 图像在夜间环境输入网络时会导致注视区域分类效果下降, 且眼镜移除部分增加了模型训练时间. Cheng 等[32]在注视分类领域首次提出混合 Transformer 模型 GazeTR-Hybrid, 使用 CNN 从人脸图像中提取局部特征, 并将特征矩阵馈送到 Transformer 中捕获全局关系, 表明混合模型在注视区域分类任务中的性能明显优于单一的 Transformer 模型, 但未对 Transformer 模型本身庞大的参数量和复杂度进行优化.

基于以上研究分析, 现有方法虽能很好地对注视区域进行分类, 但是仍然存在着一些缺陷, 例如: 头部姿态估计不准确、分类精度较低、模型结构复杂等问题. 本文提出一种并行分支 SE 注意力机制关注网络浅层中的通道和空间特征信息, 同时使用 DW 卷积降低模型复杂度, 并进一步将提取到的人脸特征和头部姿态特征进行融合, 输入到新模型中得到被试在模拟飞行过程中的注视区域分类结果.

2 基于多模态数据融合的飞行员注视区域分类网络

在计算机视觉领域, 一直使用 CNN 来处理图像任

务, 然而随着数据集的不断扩大和模型规模的增加, 传统的 CNN 架构逐渐显得乏力, 为了更好地处理大规模图像任务并提高模型的可扩展性, 近年来不断有研究将基于自注意力机制的 Transformer 模型应用到计算机视觉领域, 证明了 Transformer 具备出色的性能以应用在计算机视觉领域.

MobileViT 结合了 CNN 轻量高效和 Transformer 的自注意力机制及全局视野的优势. 相较于其他 Transformer 模型, MobileViT 利用了 CNN 提供的空间归纳偏置, 解决了传统 Transformer 模型过于依赖输入序列的位置信息问题, 并且拥有较快的推理速度, 常被应用在移动端和嵌入式设备上进行高效的图像处理任务. 将 MobileViT 与其他主流基线模型进行对比, MobileViT 在各种基准测试取得了最先进的性能. 由此选择 MobileViT 作为网络的主干并使用迁移学习方法[33], 具体而言, 采用了基于模型的迁移方式, 在自制数据集 FlyGaze 上以 224×224 的输入分辨率微调在 ImageNet-1K 数据集上预训练的 MobileViT-XXS 模型, 包含 36 531 张训练图像和 9 137 张测试图像. 对 MobileViT 模型的最后一个 1×1 卷积层和分类器中的全连接层进行微调, 这两个部分的微调涉及新任务的特征提取和分类部分, 以便适应新任务的特征分布和类别标签.

轻量级飞行员注视区域分类网络 PilotT 的网络结构如图 1 所示. 将大小为 224×224 的人脸图像作为网络输入, 通过卷积核为 3×3, 步长为 2 的卷积操作将图像形状改变为 112×112. Layer1 层中的 MV2 improved SE 表示在 MobileNetV2 中加入基于改进 SE 机制的逆残差块. 在骨干结构中, 通过 4 个 Layer 构建不同大小的特征图, Layer 中的 L=N 表示将 Transformer 块重复 N 次. Layer2 中存在 2 个 MV2 improved SE 块, 其余 Layer 中都存在重复堆叠的 MobileViT Block+DW Conv, 表示引入 DW 卷积的 MobileViT Block. 有关特征图在网络中的输出特征表示及维度变化见表 1.

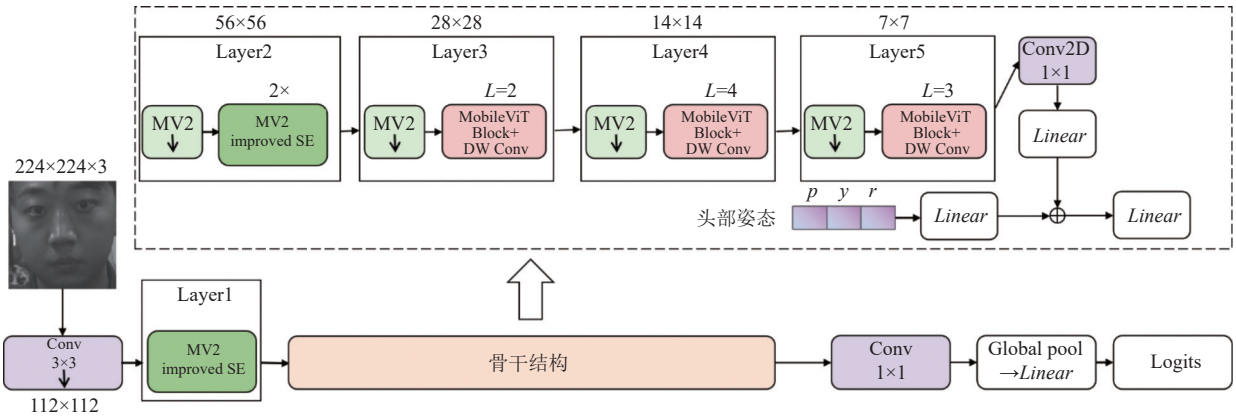

图 1 PilotT 模型的整体框架

表 1 特征维度变化

层数	输入	输出
Conv1	3×224×224	16×112×112
Layer1	16×112×112	16×112×112
Layer2	16×112×112	24×56×56
Layer3	24×56×56	48×28×28
Layer4	48×28×28	64×14×14
Layer5	64×14×14	80×7×7

2.1 多模态数据融合模块

人类在自然现象下做出的动作会表现出十分丰富的特征, 单一模态的数据很难表现出完整的特征信息[34]. 多模态数据融合将整体信息进行整合, 是提高模型分类精度的有效技术. 提出的多模态数据融合模块 MDFM 结构如图 2 所示.

俯仰角、偏航角、滚动角表示头部姿态的 3 个欧拉角, 三者堆叠为一个 3×1 张量. 由于不同类型的特征具有不同的尺寸, 因此需要对图像特征使用 1×1 卷积进行通道降维. 经实验对比, 选择在 Layer5 层之后进行特征融合操作. 在神经网络的前向传播过程中, 将多维图像特征向量(B,C,H,W), B表示批量大小, C,H,W表示图像的通道数、高度和宽度. 通过展平操作变为二维向量$(B,C\times H\times W)$. 而后使用全连接层将该二维向量映射至 100 维度的向量, 而头部姿态作为一个 3×1 的向量, 需要通过线性变换将特征向量映射至与图像

特征向量可以进行拼接的维度, 如式 (1) 所示.

$$\begin{cases} X_{\text{out}} = Linear(X_f, S_f) \\ hp_{\text{out}} = Linear(HP_f, S_f) \end{cases} \tag{1}$$

其中, $Linear$ 表示全连接层, X_f 表示图像特征向量, HP_f 表示头部姿态特征向量, S_f 为可调整的映射维度, X_{out}、hp_{out} 分别表示图像特征和头部姿态特征经过全连接层映射后的向量. 由于文中所使用的 2 个张量具有相似的特征表示, 使用加法操作更容易进行梯度传播, 如式 (2) 所示.

$$X = X_{\text{out}} + hp_{\text{out}} \tag{2}$$

最终使用一个全连接层将拼接后的向量 X 映射回初始维度, 然后将其展平并输入分类层.

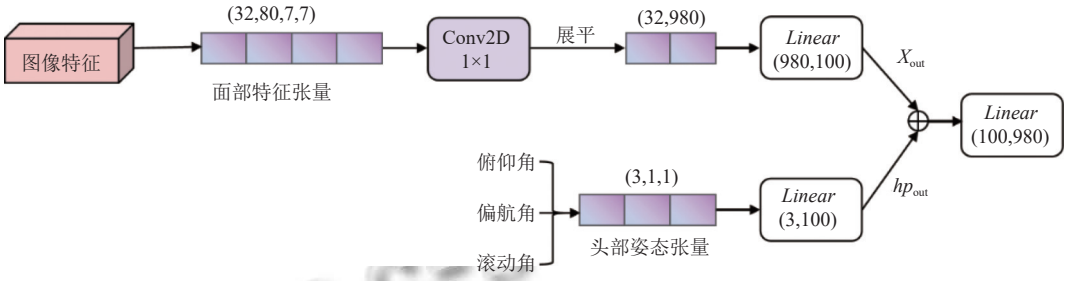

图 2 MDFM 结构图

2.2 基于并行分支 SE 机制的逆残差块

由于 SE 机制缺乏对空间维度上重要特征的关注, 为解决这一限制, 设计了一种基于并行分支 SE 注意力机制的逆残差块, 如图 3 所示. 引入 2 个并行分支来分别计算通道注意力和空间注意力. 在通道注意力分支中, 特征 $X \in R^{H\times W\times C}$ (R 表示真实数的集合) 通过全局平均池化和全局最大池化处理后分别获得 $C_{\text{avg}} \in R^{1\times 1\times C}$ 和 $C_{\max} \in R^{1\times 1\times C}$, 将这两者连接起来以获得 $C_{\text{con}} \in R^{1\times 1\times 2C}$, 然后通过 1×1 卷积层和 Sigmoid 激活函数处理后获得通道权重 $C_{\text{con}} \in R^{1\times 1\times C}$, 最终将该权重与原始特征图相乘以恢复原始大小 $X_C \in R^{H\times W\times C}$.

在空间注意力分支中, 特征 $X \in R^{H\times W\times C}$ 首先通过 1×1 卷积层, 随即输入 Sigmoid 函数将输入值映射到 0–1 之间的范围, 表示对应特征的重要程度, 得到空间权重 $X_s \in R^{H\times W}$, 最终与原始特征图相乘以恢复原始大小 $X_S \in R^{H\times W\times C}$, 这种机制有助于模型聚焦更有用的特征, 同时抑制无关特征, 从而提高模型的鲁棒性. 最终将 2 个分支得到的特征 X_C 和 X_S 相加得到最终的结果.

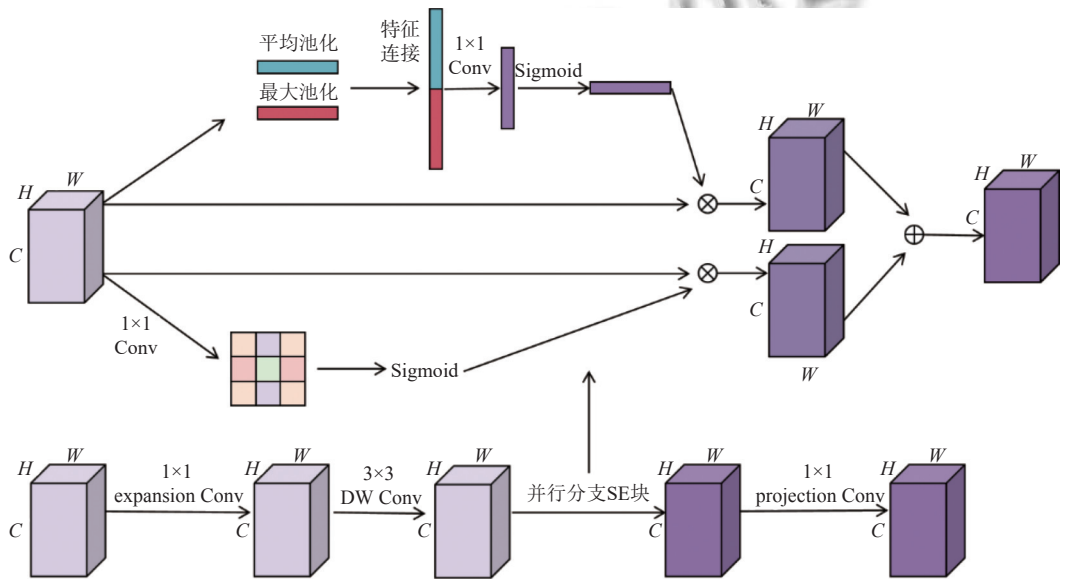

图 3 基于并行分支 SE 的逆残差块

并行分支 SE 块可以从通道和空间两个维度上增强模型对特征的注意力, 虽然额外增加了一条空间注意力分支, 但由于使用了简单的1×1卷积和 Sigmoid 激活函数, 额外的计算开销相对较低, 使得改进后的模型在保持性能的同时提高分类准确率.

设计的逆残差块公式表达如式 (3), 其中$Norm(x)$对输入特征x进行归一化操作, $DWConv$对归一化的特征进行 DW 卷积操作, SE_p表示学习输入特征图的空间及通道权重, $Proj$将经过处理的特征投影回原始维度, 有助于保留高层次的语义信息.

$$MV2(x) = x + Proj(SE_p(DWConv(Norm(x)))) \quad (3)$$

式 (3) 将原始输入特征与经过处理的特征相加, 实现残差连接, 有效减轻网络内信息丢失问题.

2.3 改进的 MobileViT Block

该结构由 3 个子模块组成: 局部特征编码、全局特征编码和特征融合模块. 首先通过局部表示模块提取输入图像局部特征. 其次将特征张量投影到高维空间并送入全局表示模块, 使用 Transformer 来扩展感受野, 以全局处理代替局部处理.

进行像素级的注意力计算时, 若每个像素都关注相邻的像素, 会造成计算资源的浪费, 相邻像素在分辨率较高的特征图上收益相对较低, 而增加的计算成本远远超过在准确性上的增益. 当特征图的高度、宽度和通道数分别为H,W,C, 则原始计算成本如式 (4) 所示.

$$Cost = \mathrm{O}(HWC) \quad (4)$$

自注意力计算成本如式 (5) 所示, 即理论上的计算成本只有原来的 1/4.

$$Cost_{\mathrm{att}} = \mathrm{O}(HWC/4) \quad (5)$$

利用展开和折叠操作将数据调整成适用于自注意力计算的格式. 如图 4 所示, 将相同颜色的色块展平到一个序列中, 并行计算每个序列, 再将这些序列折叠回原特征图的形状. 最终通过一个 1×1 卷积层将通道数调整为原始大小, 获得与原始输入X具有相同维度的$\widetilde{X} \in R^{H\times W\times C}$, 当步长为 1 时, 应用捷径分支, 将全局特征和局部特征沿通道方向拼接以获得$\widetilde{X} \in R^{H\times W\times 2C}$. 最终, 得到通道数为$2C$的特征图后, 使用 DW 卷积进行特征融合, 得到输出$Y \in R^{H\times W\times C}$, 改进的 MobileViT Block 的结构如图 5 所示.

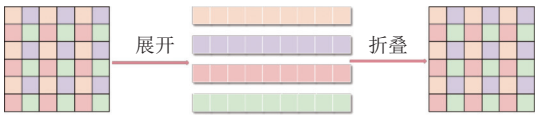

图 4 展开和折叠

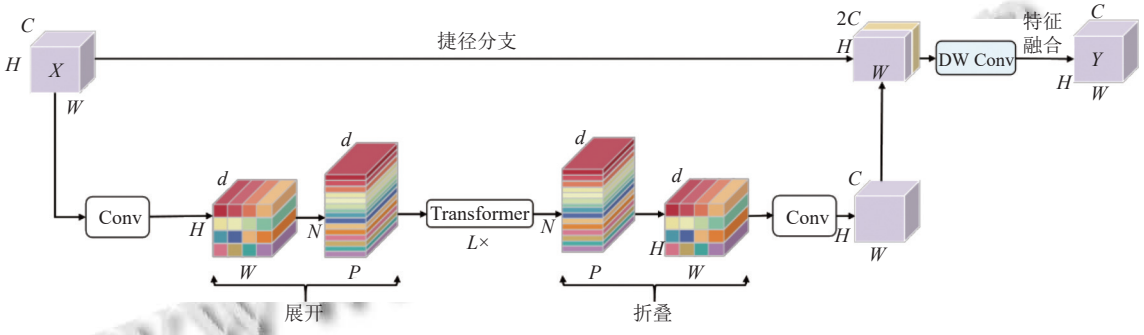

图 5 加入 DW 卷积的 MobileViT Block

DW 卷积包含逐通道卷积和逐点卷积. 逐通道卷积的参数量计算见式 (6), 计算量见式 (7).

$$F_{1\_\mathrm{num}} = F_w \times F_h \times C \quad (6)$$

$$F_{1\_\mathrm{cal}} = F_w \times F_h \times (H - F_h + 1) \times (W - F_w + 1) \times C \quad (7)$$

其中, $F_w \times F_h$表示卷积核大小, C为输入图像的通道数, H、W为输入图像的大小. 经过逐通道卷积后的特征图数量与输入层的通道数相同, 无法扩展特征图, 为有效利用不同通道在相同空间位置上的特征信息, 需要使用逐点卷积. 逐点卷积的卷积核大小为$1\times 1\times M$, M为上一层的通道数, 将上一层特征在深度方向上进行加权组合, 生成新的特征图. 逐点卷积的参数量计算见式 (8), 计算量见式 (9).

$$F_{2\_\mathrm{num}} = 1 \times 1 \times C \times M \quad (8)$$

$$F_{2\_\mathrm{cal}} = 1 \times 1 \times P_w \times P_h \times C \times M \quad (9)$$

其中, P_w和P_h表示经过逐通道卷积得到的输出特征的大小.

而普通卷积的参数量计算公式和计算量公式如式 (10)、式 (11) 所示:

$$F_{3\_num} = F_w \times F_h \times C \times M \tag{10}$$

$$F_{3\_cal} = F_w \times F_h \times (H - F_h + 1) \times (W - F_w + 1) \times C \times M \tag{11}$$

DW 卷积的参数量和计算量只有标准卷积的 $1/n$, n表示输入通道的数量, 表明 DW 卷积相较于传统卷积可以有效降低模型的复杂度.

3 实验设计

以模拟飞行任务为背景, 研究被试在飞行任务中的视线方向并对注视区域进行准确分类. 如图 6 所示, 点A表示被试头部位置及视线方向, 在双目相机各成一像, 记作A_L, A_R. 参数(p, y, r)分别代表头部的 3 个方向, 对应俯仰角 (pitch)、偏航角 (yaw) 和滚动角 (roll). 此外, (x, y, z)表示面部中心相对于相机中心的位置, (x, y)是视线与屏幕相交的注视点.

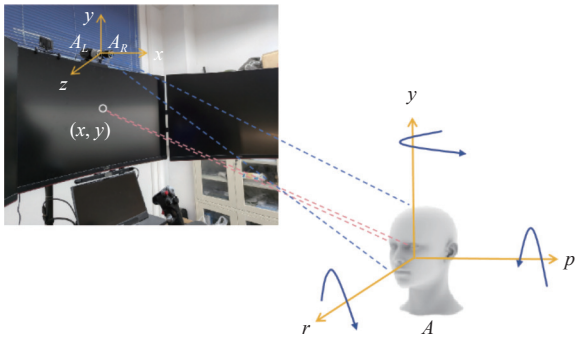

图 6 视线跟踪系统

3.1 实验方法

采用六轴模拟飞行平台作为模拟飞行环境. 平台配备三屏拼接显示器、飞行座椅、HOTAS 摇杆和脚舵, 如图 7 所示. 并额外为平台安装了一部由两个水平放置的灰度相机组成的双目相机, 用来采集被试在执行飞行任务期间的人脸图像, 同时安装一个近红外补光灯, 确保相机成像质量.

使用 DCS World 模拟飞行软件, 模拟机型为 Su-25T 战斗机. 模拟飞行过程中, 根据被试的注视区域分布情况, 将飞机左视窗、平视显示窗和右视窗作为观察舱外状况的 3 个注视区域. 将空速表、航向表和起落架作为关注机身状况的 3 个注视区域, 如图 8 所示. 这 6 个不同的注视区域代表了可以保持安全飞行的关键区域, 可以全面了解被试在飞行过程中的视线分布情况.

图 7 六轴模拟飞行平台

图 8 飞行座舱注视区域划分

舱内区域标为 1、2、3 号, 对应着起落架、空速表和航向表. 舱外区域标为 0、4、5 号, 对应平视显示窗、左视窗和右视窗. 每个区域在实际操作中具有不同的功能, 如表 2 所示.

表 2 区域功能说明

编号	名称	作用	重要性
1	起落架	为飞机起降提供支持	飞行安全的关键组件, 需要精准操作
2	空速表	测量机身相对于空气的速度	了解飞机性能和调整空速的关键仪器
3	航向表	显示飞机方向, 提供导航	确保飞机沿预定航线飞行的关键导航工具
0	平视显示窗	投影飞行信息至飞行员视野, 提升信息获取效率	提高飞行员对飞行数据的感知能力
4	左视窗	提供左侧环境视野	增强飞行员对飞机左外侧空间的感知
5	右视窗	提供右侧环境视野	增强飞行员对飞机右外侧空间的感知

飞行员对机舱内外信息的认知依附于具体的飞行任务, 由此选用五边飞行任务. 五边飞行作为飞行员必须要掌握的任务, 包含了大多数型号飞机在完成任意航线中会经历的各个阶段, 如图 9 所示. 要求被试在起飞、巡航和着陆阶段保持正确的高度和航向以及在正确时机打开降速伞、收放起落架等具体任务, 任务要求如表 3 所示.

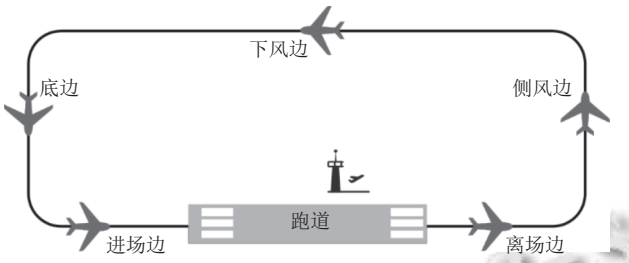

图 9 五边飞行实验航线

表 3 任务要求

阶段	航线	任务要求
起飞	离场边	航向95°, 离地时收回起落架, 爬升至800 m高度
巡航	侧风边	保持800 m左右高度平飞, 根据轨迹烟雾左转至5°航向
	下风边	保持800 m左右高度平飞, 根据轨迹烟雾左转至275°航向
	底边	保持800 m左右高度平飞, 根据轨迹烟雾左转至185°航向
降落	进场边	转向跑道方向95°, 离地200 m打开降速伞, 离地50 m放下起落架

3.2 数据采集及处理

3.2.1 FlyGaze 数据集

招募 12 名 (2 名女性, 10 名男性, 22–28 岁) 视力正常或经过视力矫正的志愿者参与实验, 被试提前学习基础飞行课程, 正式实验开始前均可熟练使用六轴模拟飞行平台进行完整的五边飞行任务. 为避免干扰, 在准备好实验环境后, 仅留一名被试单独完成实验.

一次完整的五边飞行实验根据不同被试的驾驶习惯大约持续 23–40 min, 采集程序同步采集模拟飞行过程中被试注视目标时的人脸图像. 为保证人脸图像质量, 采集程序会通过判断自动丢弃模糊或不完整的人脸图像. 6 种类型注视区域的人脸图像数据集如表 4 所示, 每张图像都有对应的头部姿态数据, 共计 45 668 张图像. 为确保数据的随机性, 使用随机拆分策略将整体数据集的 80% 作为训练集, 由 36 531 张图像组成, 剩下的 20% 作为测试集, 由 9 137 张图像组成.

3.2.2 图像处理

为解决相机采集图像时的畸变问题, 采用张正友标定法[35]对畸变图像进行矫正. 使用打印的棋盘格作为标定板, 通过棋盘格的角点来计算相机的内外参数. 不同角度的标定图像如图 10 所示.

表 4 数据集构成

座舱	目标区域	数据集	训练集	测试集
舱内区域	1-起落架	7 532	6 025	1 507
	2-空速表	7 471	5 976	1 495
	3-航向表	7 218	5 774	1 444
舱外区域	0-平视显示窗	7 931	6 344	1 587
	4-左视窗	7 691	6 152	1 539
	5-右视窗	7 825	6 260	1 565

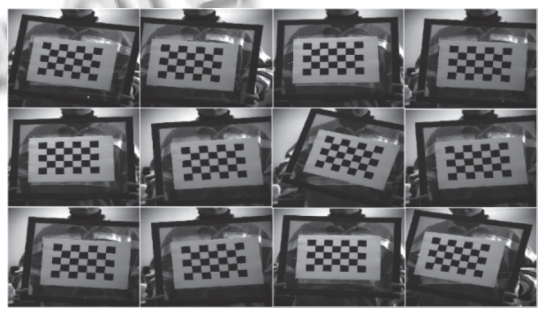

图 10 棋盘格标定图

使用 Matlab 中的 Stereo Camera Calibrator 程序进行标定工作. 传入由双目相机拍摄的 20 组包含不同角度的大小为 30 mm 的棋盘格图像. 在标定过程中, 使用重投影误差作为标定结果的评价标准, 重投影误差即投影的点与图像上的测量点的误差像素. 如图 11 所示, 筛除重投影误差较大的标定板, 总体平均误差为 0.14 个像素, 选取 26 张有效的标定图像. 图 12 以 3D 图的形式显示两个相机和校准平面的空间关系.

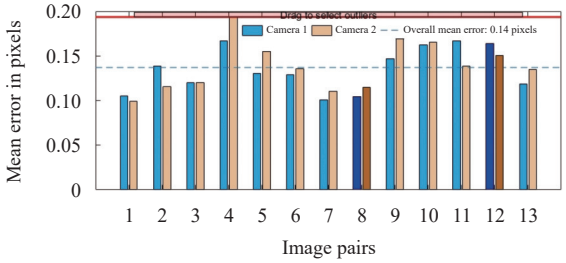

图 11 重投影误差

完成标定后, 获取到双目相机的内参矩阵、畸变系数 (distortion coefficient, D), 以及两个相机相对的旋转矩阵和平移矩阵. D是由径向畸变系数 (radial distortion, R) 和切向畸变系数 (tangential distortion, T) 组成, R、T以及D可由式 (12) 表示, 校正前后的对比图像如

图 13 所示.

$$\begin{cases} R=(r_1,r_2,r_3) \\ Z=(z_1,z_2,z_3) \\ D=(r_1,r_2,z_1,z_2,r_3) \end{cases} \tag{12}$$

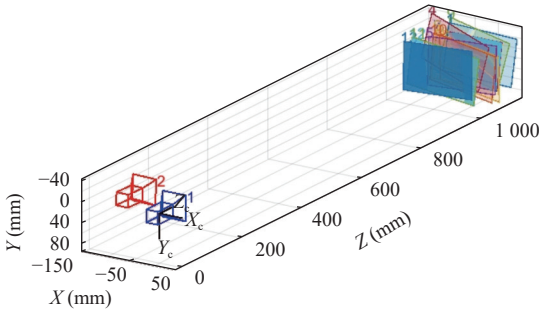

图 12 三维空间关系

(a) 校正前 (b) 校正后

图 13 校正前后人脸图像对比

Bazarevsky 等[36]提出一种高效的轻量级人脸检测框架 BlazeFace, 该框架使用改进的 MobileNet 作为特征提取器, 引入能够高效利用 GPU 的锚点 (anchor) 机制. BlazeFace 人脸检测器在近距离正脸场景下的人脸检测任务中兼具准确性和快速推理的优势. 采用 BlazeFace 人脸检测器对经过校正后的 2D 人脸图像进行分割和提取, 结果如图 14 所示.

图 14 提取后的人脸图像

3.2.3 头部姿态解算

头部姿态估计是指在数字图像里通过相关流程将以像素为基础的头部表示转换为方向的过程. 本文采用基于图像的头部姿态估计方法[37], 与其他面部视觉任务相同, 基于图像的头部姿态估计必须在面对各种图像变化因素时表现出不变性, 例如镜头畸变引起的相机失真、三维场景投影到二维图像平面的投影几何问题[38]. 为了确保头部姿态估计的准确性, 利用第 3.2.2 节获得的实验参数, 建立准确的相机参数模型, 定义一个具有 5 个关键点的三维面部模型进行二维人脸关键点检测及匹配. 最终通过旋转矩阵解算头部姿态.

dlib-Python 库提供的人脸检测器可以确定图像中人脸的具体位置, 而关键点检测器可以检测人脸上的重要特征点, 如眼角、鼻子和嘴角等. 借助该库来确定图像中人脸位置和人脸的重要特征点, 人脸特征点检测结果如图 15 所示.

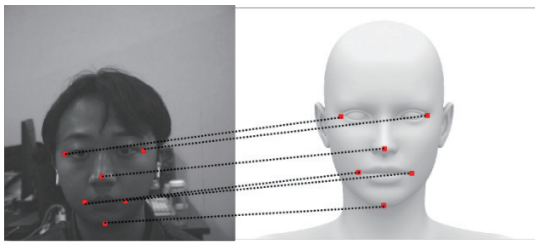

图 15 二维和三维特征点

通过相机标定和 dlib-Python 库的人脸检测与关键点检测, 可以计算人脸的关键特征点的位置和空间关系, 推断头部的旋转和倾斜角度, 从而获得精确和稳定的头部姿态估计结果, 有效补偿视线偏移带来的注视区域估计误差. 头部姿态估计结果如图 16 所示.

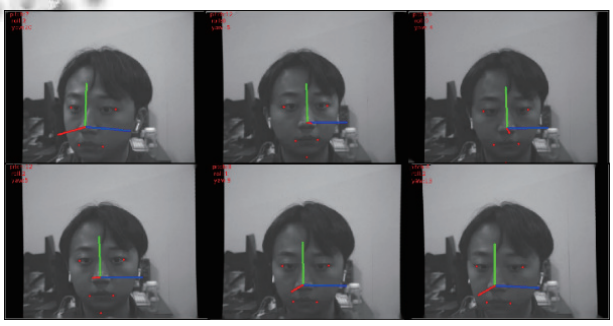

图 16 头部姿态估计结果

4 实验结果与分析

4.1 准确率对比

使用 Ubuntu 20.04 操作系统、PyTorch 1.10.0 深度学习框架、Python 3.8 编程语言. 使用 CUDA 11.3

和 cuDNN v8.2.1 加速模型进行训练.

CPU 为 Intel Xeon E5, 内存 128 GB. GPU 为 NVIDIA TITAN V×4, 显存 48 GB.

为评估不同模型对注视区域分类效果的影响, 将 PilotT 网络模型与现有主流网络模型进行对比, 微调网络进行训练, 配置相同的训练参数. 对自制数据集进行训练, 训练轮次设置为 100, 批量大小设置为 32, 使用交叉熵损失函数, 初始学习率设置为 0.001, 使用余弦退火算法调整学习率, 每次训练结束时保存最优网络权重模型. 将网络模型的输出值与真实标签值进行对比, 分析当前模型确定的区域是否和真实标签值一致作为注视区域分类模型的评估指标. 对比实验结果如表 5 所示.

表 5 不同模型分类结果比较

网络结构	图像分辨率	人脸图像 Top-1 Acc (%)	人脸图像+头部姿态 Top-1 Acc (%)
MobileNetV2	224×224	86.56	87.64
SwinT	224×224	88.17	89.04
ViT-B/16	224×224	79.34	82.35
MobileViT-XXS	224×224	89.46	90.58
ResNet50	224×224	88.25	90.03
EfficientNet	224×224	82.97	84.79
ConvNeXt	224×224	89.68	90.89
MobileNetV3	224×224	87.69	89.51
PilotT	**224×224**	**90.63**	**92.79**

由表 5 可知, MobileNetV2/V3、ResNet50、EfficientNet 以及 ConvNeXt 这些传统 CNN 网络模型均表现出稳定的注视区域分类效果. 其中 ConvNeXt 在 2 种场景下的分类效果为 CNN 模型中的最优, 分别达到 89.68% 和 90.89%. EfficientNet 的分类效果均低于其他的 CNN 模型, 和 ConvNeXt 相差 6.71% 和 6.1%. MobileNetV3 相较于 MobileNetV2 在 2 个场景下的准确率分别提升 1.13% 和 1.87%, 这是由于前者通过 NetAdapt 算法获取卷积核和通道的最佳数量, 并引入了 SE 机制.

SwinT、ViT-B/16、MobileViT-XXS 作为 Transformer 模型的代表, 之间展现出较大的差异. ViT-B/16 在 2 个场景下的分类准确率均为最后一名, 与 Transformer 模型中排名第 1 的 MobileViT-XXS 分别相差 10.12% 和 8.23%. SwinT 和 ResNet50 表现接近, 仅有 0.08% 和 0.99% 的差异.

为进一步验证 PilotT 模型的有效性, 将 9 个网络在相同飞行任务下对 6 个注视区域的 Top-1 分类准确率进行详细比较. 如图 17 所示, 其中横轴表示划分好的 6 个注视区域, 纵轴表示 Top-1 分类准确率, 图 17(a) 表示仅将人脸图像作为输入时的情况, 图 17(b) 表示将融合后的数据作为输入的情况.

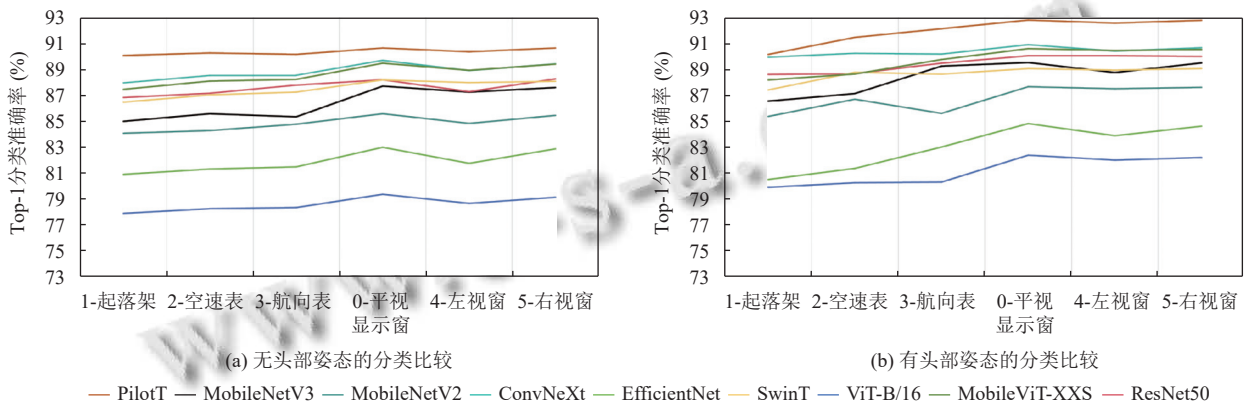

图 17 注视区域分类结果

在不包含头部姿态时, ViT-B/16 和 EfficientNet 的分类准确率明显低于其他模型, 尤其对区域 1 (起落架)、区域 2 (空速表) 和区域 3 (航向表) 的分类效果与其他模型相差较大, 说明 ViT-B/16 和 EfficientNet 对类间差异性较小的人脸图像的识别精度较差. 相比于 MobileNetV2, MobileNetV3 对各个区域的分类准确率都有不同程度的提升, 但与 ResNet50 相比, 除了对区域 4 (左视窗) 的分类效果优于 ResNet50, 其他区域准确率均低于 ResNet50, 这是由于 MobileNetV3 原有轻量级结构无法获取丰富的特征表示, 网络表征能力较差. SwinT 在面积较小区域 (区域 1、2、3) 上的分类效果不如 ResNet50, 在面积较大区域, 如区域 0 (平视显示窗)、区域 4 (左视窗)、区域 5 (右视窗) 上的分类效果与 ResNet50 相似, 表明区域面积的大小会影响模型的分

类决策. ConvNeXt 在各区域的指标均优于 ResNet50, 但由于引入了块链接结构, 网络泛化能力较差. 与 MobileViT-XXS 相比, PilotT 在区域 1、2、3、4 上的分类准确率提升最大, 分别提升了 2.61%、2.17%、1.93% 和 1.73%, 原因在于 PilotT 模型能同时捕捉特征图的通道和空间信息, 并在全局范围内学习特征之间的长距离依赖关系, 充分结合各阶段的特征信息, 从而提升网络表征能力.

实验表明在结合人脸图像和头部姿态的场景下, 所有模型的分类准确率都有不同程度的提升, 验证了头部姿态在补偿视线偏移、提升模型分类准确率方面的作用. 区域 0、4、5 的分类效果最好, 原因在于双目相机安装在飞行员正前方, 当飞行员注视这些区域时, 眼部和头部运动变化差异明显, 相较于其他区域更加容易区分. PilotT 网络对区域 1、2、3 进行分类时最易产生混淆, 通过分析发现这 3 个注视区域在机舱仪表盘上距离相隔较近, 飞行员在注视这些区域时视线方向和头部朝向均偏向左方, 特征提取网络提取出的人脸特征和头部姿态特征比较相似, 导致这些区域分类的准确率较低. 这也支持了以下假设: 对于从一个区域到另一个区域, 特别是对于具有较远物理距离的区域 (例如: 区域 4 和 5), 分类准确率增加.

4.2 飞行可视化分析

为验证 PilotT 模型在实时模拟飞行任务中的有效性和性能, 使用 PyTorch 中的 Thop 库计算了模型的浮点运算次数和参数量. 统计结果如表 6 所示.

表 6 不同模型性能对比

模型	GFLOPs	参数量 (M)
MobileNetV3	0.29	5.48
SwinT	4.37	28.26
ConvNeXt	4.46	28.56
MobileViT	0.27	1.27
PilotT	0.28	1.32

通过表 6 可知, PilotT 模型的参数量和浮点运算次数明显小于其他模型, 与原模型相比, 本文所做的改进能在提高模型分类精度的情况下保持同样少的模型计算量, 使其能够应用在对实时性要求较高的场合.

通过完成实时的飞行任务对使用 PilotT 模型收集到的注视数据进行可视化分析, 可以判断飞行学员在任务中的操作是否正确, 并及时纠正错误的操作以提升飞行训练效果. 以下设定 2 个不同的飞行任务, 每个任务随机挑选一名被试执行.

4.2.1 左转弯任务

PilotT 模型收集了被试从航向 15°转到航向 285° 的左转弯任务中的注视数据, 具体数据见表 7.

表 7 左转弯任务中注视时长及注视次数统计结果

注视区域	注视时长 (ms)	注视次数
1-起落架	243	12
2-空速表	25 341	1 521
3-航向表	63 275	5 624
0-平视显示窗	617	52
4-左视窗	18 347	935
5-右视窗	198	15
其他区域	324	27

总的注视时长为 108 345 ms, 总注视次数为 8 186 次. 由表 7 可知飞行学员在执行左转弯任务时的主要注视区域为航向表和空速表, 其次为左视窗, 其余区域的注视次数相对较少, 其他区域可能由扫视行为造成. 可以推断, 在执行该任务时飞行学员能够通过航向表准确掌握飞行方向并寻找目标航向 285°; 同时, 通过空速表将飞机速度控制在合理范围内, 以避免因速度过快导致航向丢失或失速撞击等问题. 标准的左转向任务并不需要过多关注左视窗, 但该学员下意识地给予左视窗过多的关注, 评判结果为操作不合格, 后续需要纠正不规范行为.

4.2.2 起飞任务

被试在起跑道上保持 15°航向, 空速达到 170–190 km/h 时开始起飞, 当离地高度大于 10 m 时收起起落架, 完成起飞任务. 该过程中由 PilotT 收集到的注视数据如表 8 所示.

表 8 起飞任务中注视时长及注视次数统计结果

注视目标	注视时长 (ms)	注视次数
1-起落架	6 276	520
2-空速表	35 279	3 108
3-航向表	19 723	1 254
0-平视显示窗	8 814	893
4-左视窗	145	12
5-右视窗	156	12
其他区域	213	16

总的注视时长 70 606 ms, 总注视次数为 5 815 次. 由表 8 可知飞行学员的主要注视区域为航向表和空速表, 其次注视的区域为平视显示视窗和起落架, 注视相对较少的区域为左视窗和右视窗, 其他区域可能由扫视行为造成. 通过分析可知, 该学员时刻关注空速表以判断飞机是否达到成功起飞的标准; 同时通过航向表

判断当前飞机是否偏离航向; 当飞机离开地面后, 通过起落架指示仪表判断是否应该收起起落架, 并确认起落架是否正常工作. 该学员的操作十分标准, 评判结果为操作合格.

使用 PilotT 模型对飞行训练任务中飞行学员的注视区域进行分类, 收集并分析训练数据对增强飞行训练效果、提高实际飞行操作的安全性和效率具有重要意义.

4.3 消融实验

为验证改进后模型的收敛性, 对比了改进前后模型在 FlyGaze 数据集上每轮训练的损失值, 如图 18 所示. 可以看出在经过 100 个 epoch 后, 2 个模型的损失值均趋于稳定, 在第 62 轮后损失差距越来越明显. 最终改进后模型的损失值为 0.502, 原始模型的损失值为 0.533. 实验结果表明改进模型在训练过程中可以更快地收敛.

同时为验证 PilotT 模型各改进模块对注视区域分类效果的影响, 进行了消融实验, 对 PilotT、MobileViT+MDFM、MobileViT+improved SE 以及 MovbileViT 共计 4 个模型进行对比实验, 实验结果如图 19 所示.

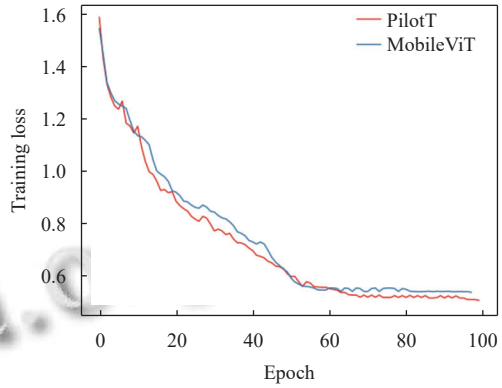

图 18 改进模型前后训练损失图

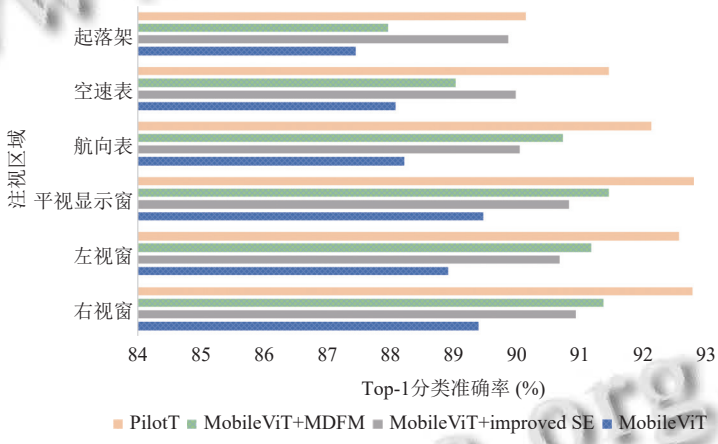

图 19 消融实验结果对比

(1) MobileViT+ MDFM. 为验证 MDFM 模块对补偿头部偏转造成的视线方向误差的影响, 对比了在没有该模块下的效果. 从图 19 中可见, 相较于 MobileViT, 区域 0–5 的准确率分别提升了 1.99%、0.52%、0.95%、2.51%、2.26%、1.97%. 由此证明该模块对于提升网络分类准确率的有效性.

(2) MobileViT+improved SE. 在 MobileViT 的基础上加入基于改进的并行分支 SE 机制的逆残差块, 对比了在没有该机制下的分类效果. 在缺少对网络浅层通道和空间进行特征提取的操作时, 由于飞行员左方 3 个注视区域 (区域 1、2、3) 的特征信息较为相似, 导致模型无法有效地区分这 3 个区域之间的特征差别, 从而对此片区域的分类效果较差. 而加入 improved SE 机制的模型显著提高了类内差距较小的人脸细粒度图像的分类准确率. 由此可以证明该改进模块在 PilotT 模型中的有效性.

5 结束语

设计一个兼具高性能和高准确率的视线分类模型对规范飞行学员的操作以增强飞行训练效果, 以及对飞行员进行疲劳监测和意图识别具有实际应用价值.

本文提出了一种基于多模态数据融合的飞行员注视区域分类网络 PilotT, 加入 DW 卷积代替结构中的普通卷积减小模型复杂度, 设计并引入了一种基于并行分支 SE 机制的逆残差块, 同时提取网络浅层的通道和空间特征信息, 并与全局特征之间建立长距离依赖

关系. 同时设计并引入了多模态数据融合模块, 充分融合人脸特征与头部姿态特征. 在自制数据集 FlyGaze 与其他模型进行实验对比, 实验结果表明 PilotT 在保持最小参数量和浮点运算次数的同时具有最高的分类准确率.

参考文献

1 苟超, 卓莹, 王康, 等. 眼动跟踪研究进展与展望. 自动化学报, 2022, 48(5): 1173–1192.

2 Marvasti-Zadeh SM, Cheng L, Ghanei-Yakhdan H, *et al*. Deep learning for visual tracking: A comprehensive survey. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(5): 3943–3968. [doi: 10.1109/TITS.2020.3046478]

3 Jiao LC, Wang D, Bai YD, *et al*. Deep learning in visual tracking: A review. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(9): 5497–5516. [doi: 10.1109/TNNLS.2021.3136907]

4 Akinyelu AA, Blignaut P. Convolutional neural network-based methods for eye gaze estimation: A survey. IEEE Access, 2020, 8: 142581–142605. [doi: 10.1109/ACCESS.2020.3013540]

5 Jiang JQ, Zhou XL, Chan S, *et al*. Appearance-based gaze tracking: A brief review. Proceedings of the 12th International Conference on Intelligent Robotics and Applications. Shenyang: Springer, 2019. 629–640.

6 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

7 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.

8 Sifre L, Mallat S. Rigid-motion scattering for texture classification. arXiv:1403.1687, 2014.

9 Howard A, Sandler M, Chen B, *et al*. Searching for MobileNetV3. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1314–1324.

10 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.

11 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

12 Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision Transformer. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

13 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

14 Tan MX, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 6105–6114.

15 Liu Z, Mao HZ, Wu CY, *et al*. A ConvNet for the 2020s. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11966–11976.

16 Aunsri N, Rattarom S. Novel eye-based features for head pose-free gaze estimation with Web camera: New model and low-cost device. Ain Shams Engineering Journal, 2022, 13(5): 101731. [doi: 10.1016/j.asej.2022.101731]

17 Lundgren M, Hammarstrand L, McKelvey T. Driver-gaze zone estimation using Bayesian filtering and Gaussian processes. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(10): 2739–2750. [doi: 10.1109/TITS.2016.2526050]

18 Wang YF, Yuan GL, Mi ZT, *et al*. Continuous driver's gaze zone estimation using RGB-D camera. Sensors, 2019, 19(6): 1287. [doi: 10.3390/s19061287]

19 Cheng YH, Zhang XC, Lu F, *et al*. Gaze estimation by exploring two-eye asymmetry. IEEE Transactions on Image Processing, 2020, 29: 5259–5272. [doi: 10.1109/TIP.2020.2982828]

20 Zhang XC, Sugano Y, Fritz M, *et al*. It's written all over your face: Full-face appearance-based gaze estimation. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 2299–2308.

21 Zhang XC, Park S, Beeler T, *et al*. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 365–381.

22 Liu S, Liu DP, Wu HY. Gaze estimation with multi-scale channel and spatial attention. Proceedings of the 9th International Conference on Computing and Pattern

Recognition. Xiamen: ACM, 2020. 303–309.
23 Balim H, Park S, Wang X, *et al*. EFE: End-to-end frame-to-gaze estimation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE, 2023. 2688–2697.
24 Dai LH, Liu JG, Ju ZJ. Binocular feature fusion and spatial attention mechanism based gaze tracking. IEEE Transactions on Human-machine Systems, 2022, 52(2): 302–311. [doi: 10.1109/THMS.2022.3145097]
25 Cazzato D, Leo M, Distante C, *et al*. When I look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking. Sensors, 2020, 20(13): 3739. [doi: 10.3390/s20133739]
26 戴忠东, 任敏华. 基于表观的归一化坐标系分类视线估计方法. 计算机工程, 2022, 48(2): 230–236.
27 闫秋女, 张伟伟. 基于多模态特征融合的驾驶员注视区域估计. 计算机与数字工程, 2022, 50(10): 2217–2222. [doi: 10.3969/j.issn.1672-9722.2022.10.018]
28 Ghosh S, Dhall A, Sharma G, *et al*. Speak2Label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021. 2896–2905.
29 Ali A, Kim YG. Deep fusion for 3D gaze estimation from natural face images using multi-stream CNNs. IEEE Access, 2020, 8: 69212–69221. [doi: 10.1109/ACCESS.2020.2986815]
30 杨易蓉. 基于头眼特征融合的驾驶员视线区域估计及驾驶场景关联方法研究 [硕士学位论文]. 济南: 山东大学, 2022.
31 张名芳, 李桂林, 吴初娜, 等. 基于轻量型空间特征编码网络的驾驶人注视区域估计算法. 清华大学学报 (自然科学版), 2024, 64(1): 44–54.
32 Cheng YH, Lu F. Gaze estimation using Transformer. Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Montreal: IEEE, 2022. 3341–3347.
33 Peng YJ, Xu Y, Shi J, *et al*. Wild mushroom classification based on improved MobileViT deep learning. Applied Sciences, 2023, 13(8): 4680. [doi: 10.3390/app13084680]
34 任泽裕, 王振超, 柯尊旺, 等. 多模态数据融合综述. 计算机工程与应用, 2021, 57(18): 49–64. [doi: 10.3778/j.issn.1002-8331.2104-0237]
35 Zhang Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330–1334. [doi: 10.1109/34.888718]
36 Bazarevsky V, Kartynnik Y, Vakunov A, *et al*. BlazeFace: Sub-millisecond neural face detection on mobile GPUs. arXiv:1907.05047, 2019.
37 Murphy-Chutorian E, Trivedi MM. Head pose estimation in computer vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(4): 607–626. [doi: 10.1109/TPAMI.2008.106]
38 吴其右. 头眼协调运动目标快速捕获研究 [硕士学位论文]. 西安: 西安工业大学, 2022.

(校对责编: 张重毅)