武汉大学学报(理学版)
J. Wuhan Univ. (Nat. Sci. Ed.) 1~16

DOI: 10.14188/j.1671-8836.2025.0191

基于五官旋转位移关系表征的高效头部姿态估计

刘秋霞[1],余海洋[1],刘志锋[1],夏浩庭[2],刘婷婷[3],郭振辉[1],刘　海[1,2†]

1. 中国人民解放军南部战区总医院 内科重症医学科,广东 广州 510010;
2. 华中师范大学 人工智能教育学部,湖北 武汉 430079;
3. 香港大学 教育学部,中国香港 999077

收稿日期:2025-10-14 † 通信联系人 E-mail:hailiu0204@ccnu.edu.cn

基金项目:国家自然科学基金(62577020);湖北省自然科学基金(2025AFD621);广东省基础与应用研究基金(2025A1515010266);华中师范大学中央高校基本科研业务费项目(CCNU25ai012)

第一作者:刘秋霞,女,主管护师,现从事医学图像识别、注意力识别、重症监护等方面的研究。E-mail:liuqiuxia198371@163.com

摘　要:头部姿态估计作为计算机视觉领域的关键研究任务,在人机交互、安防监控等场景中具有广泛应用前景。然而,实际的头部姿态应用面临诸多挑战,例如严重的遮挡、光照不足和极端的姿态角度。为此,本文提出一种基于五官旋转位移关系表征的头部姿态估计方法,该方法主要包含两个模块:五官关系挖掘模块和隐秘位移特征表示模块。五官关系挖掘模块通过精细建模左眼、右眼、双耳等五官区域的空间分布及相互依赖关系,显著提升模型对头部姿态关键信息的感知与理解能力;隐秘位移特征表示模块以五官旋转过程中的相对位移为核心,构建适配遮挡场景的稳健特征结构,有效突破遮挡干扰。在BIWI、AFLW2000数据集及其遮挡版本上的实验结果表明,本文方法在遮挡条件下的头部姿态估计性能优于多种现有方法,同时在无遮挡场景中准确率较高,充分验证了方法的有效性与鲁棒性。

关　键　词:头部姿态估计;五官关系挖掘;隐秘位移特征;Transformer;图像理解

中图分类号:TP391.41　　**文献标志码**:A　　**文章编号**:1671-8836(XXXX)XX-0001-16

Efficient Head Pose Estimation Based on the Representation of Facial Features Rotation Displacement Relationship

LIU Qiuxia[1], YU Haiyang[1], LIU Zhifeng[1], XIA Haoting[2], LIU Tingting[3], GUO Zhenhui[1], LIU Hai[1,2†]

1. Department of Medical Intensive Care Unit, General Hospital of Southern Theater Command of PLA, Guangzhou 510010, Guangdong, China;
2. Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, Hubei, China;
3. Faculty of Education, The University of Hong Kong, Hong Kong 999077, China

Abstract: Head pose estimation, a key research task in computer vision, has broad application prospects in human-computer interaction, security monitoring and other scenarios. However, practical applications face many challenges, such as severe occlusion, insufficient lighting and extreme posture angles. Therefore, this article proposed a head pose estimation method based on the representation of rotational displacement relationships of facial features, which mainly included two modules: Facial Relationship Mining (FRM) Module and Covert Displacement Feature Representation (CDF) Module. Specifically, the FRM Module significantly enhanced the model's ability to perceive and understand key head pose information by finely modeling the spatial distribution and interdependencies of facial regions. The CDF Module constructed a robust feature structure suitable for occluded scenarios centered on the relative displacement of facial features during rotation, effectively breaking through occlusion interference. Experimental results on BIWI, AFLW2000 datasets and their occluded versions show that the proposed method outperforms several state-of-the-art methods in occluded conditions while maintaining comparable accuracy in non-occluded scenarios, fully verifying its effectiveness and robustness.

引用格式:刘秋霞,余海洋,刘志锋,等. 基于五官旋转位移关系表征的高效头部姿态估计[J]. 武汉大学学报(理学版),XXXX,XX(XX):1-16. DOI:10.14188/j.1671-8836.2025.0191.

LIU Qiuxia, YU Haiyang, LIU Zhifeng, *et al*. Efficient Head Pose Estimation Based on the Representation of Facial Features Rotation Displacement Relationship [J]. *J Wuhan Univ (Nat Sci Ed)*, XXXX, XX(XX):1-16. DOI:10.14188/j.1671-8836.2025.0191(Ch).

Key words: head pose estimation; facial relationship mining; covert displacement feature; Transformer; image understanding

0 引 言

头部姿态估计(Head Pose Estimation,HPE)是计算机视觉中研究头部旋转角度的重要研究方向。它作为人工智能和计算机视觉领域的热点技术之一[1],能够在多个领域提供关键支撑,例如智慧教育[2]、混合现实[3]、辅助驾驶[4]和医学监控[5]等。近年来,深度学习已成为解决HPE的基础性工具。随着真实场景应用需求的不断推进,无约束场景下的估计准确性不足以及真实场景中遮挡问题的普遍性愈发凸显。

过去数十年,头部姿态估计方法已取得长足发展。总体可分为两大类:基于面部关键点检测的HPE与基于无关键点的HPE。

基于面部关键点的HPE通常需要使用三维头部模型或精确定位面部关键点,并以此为基础推断头部姿态。事实证明,人脸轮廓在头部外形中的相对位置会显著影响人类对于头部姿态的感知[6]。面部关键点与头部姿态存在天然几何关联,检测算法在复杂场景下具备较好稳定性,被广泛应用于HPE任务。正因如此,此类基于面部关键点的方法具有显著优势,它们能够直接借助已知的与人类姿态判断相关的因素进行估计[7-11]。但该类方法仍存在明显局限:高度依赖人工标注的关键点数据,标注成本高且耗时;极端姿态与遮挡场景下易出现关键点缺失、标注不准问题;姿态估计精度极易受关键点检测误差影响,误差易逐级传递放大,难以适配无约束真实场景[12-14]。

基于无关键点的HPE不需要检测面部关键点,而是直接通过图像预测头部姿态角度。现有无关键点HPE方法已从表征方式、网络结构与训练策略等维度不断优化[15-19],通过旋转矩阵表征、偏移集合策略、轻量化网络、Transformer全局建模与多尺度注意力设计等思路,直接从图像端到端回归头部姿态。该类方法无需依赖面部关键点与三维头部模型,省去人工标注与预处理流程,对训练数据噪声具备一定容忍度,泛化性较强。但其存在固有短板,缺少面部局部结构的显式约束,易丢失细粒度姿态关联信息,在低光照、面部遮挡等复杂场景下特征挖掘不足,姿态估计精度难以进一步提升[20-22]。

在真实场景中人脸均存在或多或少的面部遮挡,这些遮挡会导致关键特征点(如眼睛、鼻子、嘴巴)的部分或完全丢失,从而影响姿态判断[23],这是影响HPE准确性的重要因素之一。

Transformer架构[24-26]的注意力机制在计算机视觉领域的应用为HPE技术突破提供了新路径。目前基于深度学习的方法中,极少专门针对面部遮挡场景开展头部姿态估计研究。现有深度学习头部姿态估计方法涵盖分割建模、关键点热图回归、多任务联合学习及非线性端到端回归等多种思路[27-30],有效提升了姿态估计的鲁棒性、推理效率与全角度适配能力,模型轻量化与实用部署性也得到持续优化。但现有基于深度学习的HPE方法仍存在明显短板,多数研究仅聚焦常规无遮挡场景下的姿态估计性能,专门针对面部遮挡工况开展定向建模与优化的工作仍相对稀缺;模型缺少遮挡感知与特征补偿机制,面对口罩遮挡、外物遮挡等真实复杂场景时,易出现特征退化、角度回归偏差大的问题,难以满足无约束实际应用需求[31-33]。基于此,本文提出了基于面部五官(文中仅指代眼、眉、鼻、口,研究未涉及舌部)旋转位移关系的HPE。

本文的主要工作贡献如下:

1) 构建五官关系挖掘模块,将五官旋转位移关联规律引入头部姿态估计任务,建模眼、耳、口、鼻不同区域旋转过程中的跨部位关联规律,挖掘面部高层语义关联特征,突破传统方法仅使用单一局部特征的缺陷,为头部姿态估计提供高辨识度的结构特征支撑;

2) 设计隐秘位移表示模块提升抗干扰能力,提出了聚焦微小关键部位的隐秘位移表征模块,捕捉这类部位的细微位移与可见性变化。即便在遮挡等复杂场景下,仍能通过关键部位特征构建稳健表征结构,有效弥补五官关系失效时的特征缺失,提升估计准确性。

1 头部姿态图像特征分析

1.1 问题陈述

在无约束真实场景下,头部姿态估计任务仍存在多重现实难点。其一,光照强度差异、画面模糊、背景杂乱等多变成像环境,会大幅干扰面部特征提取质量;其二,人物头部偏转角度过大形成的极端姿态,易造成局部面部区域缺失,破坏五官空间关联信息;其三,口罩、墨镜等外物带来的面部局部遮挡问题在现实场景中十分普遍。各类复杂干扰会严重降低模型对头部关键特征的稳定捕捉能力,最终导致头部姿态预测出现明显偏差。因此,如何有效抵御多类型场景干扰,提升无约束环境下头部姿态估计的精度与鲁棒性,是当前该领域亟需攻克的核心难题。无约束场景可分为环境无约束和姿态

无约束,第一个方面主要依赖于图像质量、光线条件和背景导致的环境无约束。在真实场景中通常会出现光线不足、背景复杂或摄像头角度不理想的情况下,此时算法可能出现较大的误差;第二个方面由于姿态无约束,头部姿态偏转较大时造成面部特征的歧义性,算法无法理解面部重要特征。

1.2 重要观察

基于上述挑战,本文在大量头部图像中观察到一个显著的规律性现象:如图1所示,随头部逐渐旋转,面部五官的可见性存在变化。

定义1 旋转位移:旋转位移是以头部0°正对姿态为基准,描述头部发生旋转时,五官中单个部位(如眼睛、鼻子、耳朵、嘴巴等)自身产生的空间位置偏移与形态形变特征。头部向不同方向旋转时,因各部位在面部的解剖位置与空间占比不同,其位移表现出显著差异性:例如图1(a)和图1(c),耳部随头部左右旋转会出现明显的暴露与遮挡位移,鼻孔随头部俯仰会产生可见性转变与细微形态压缩,眼部则伴随水平偏移与眼睑开合度变化。这种单个部位的位移幅度、方向与头部旋转角度直接相关,是捕捉头部姿态变化的基础视觉特征,为后续姿态角度分析提供初始依据。

定义2 旋转位移关系:旋转位移关系是指头部旋转过程中,五官内不同部位(无论相邻或远距离,如眼睛与耳朵、鼻子与嘴巴)之间形变及位移的内在关联约束规律。它并非单一部位的独立变化,而是强调跨部位的联动性:如图1(a)和图1(b),某一部位的特定位移(如右眼水平右移),必然对应其他部位的关联性形变(如右耳暴露面积增加、口部轻微右偏)。这种关联不受部位空间距离限制,例如眼睛与鼻孔的位移联动可反映俯仰角度,耳朵与眼部的位移关联能指示偏航方向。正是这种跨部位的位移约束关系,共同构建了头部姿态的完整几何特征,成为精准判断头部旋转角度归属的核心线索。

定义3 隐秘位移感知:隐秘位移感知是针对头部姿态估计中遮挡导致五官关系表征失效的问题,提出的一种特征表示方法。其以五官中面积微小但具有姿态指示性的关键部位(如鼻孔)为感知核心,这类部位位移幅度虽小,却能直接关联特定姿态趋势:如图1(c)和图1(d)的俯仰场景中,鼻孔随低头动作由可见转为遮挡,其细微的形态压缩与可见性变化,可精准反映头部“向下看”的姿态特征。当眼、耳等大部位因遮挡无法提供有效信息时,通过捕捉这类隐秘部位的位移与状态变化,能突破遮挡干扰,实现抗遮挡的姿态特征补充,为头部姿态的精准估计提供关键且鲁棒的线索支撑。

基于上述定义,具有空间位置偏移与形态形变特征的五官部位表征为旋转位移表征,具有跨部位联动约束规律的多部位关联表征为旋转位移关系表征。二者可结合隐秘位移感知补充建模:如图1所示,头部不同旋转方向下五官各部位的独立空间变化,即为旋转位移表征,隐秘位移感知可精准捕捉其状

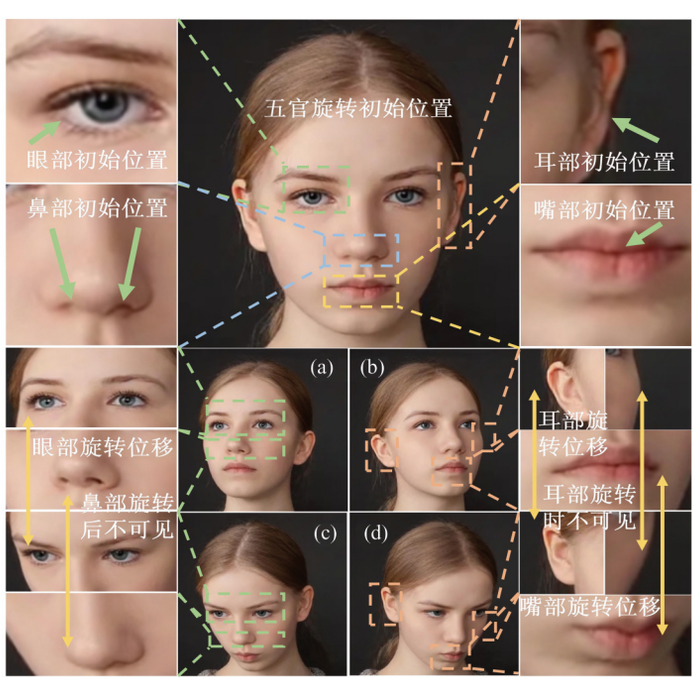


图1 五官旋转引起的面部结构可见性变化示意图

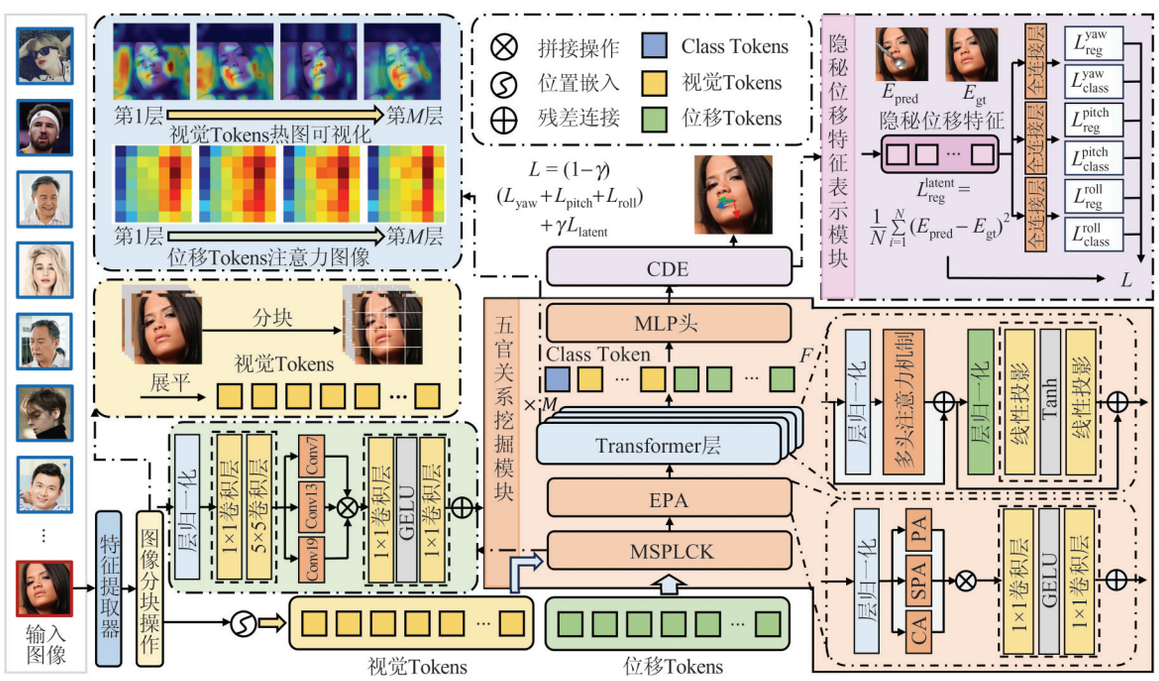
Fig. 1 Schematic diagram of changes in visibility of facial structure caused by rotation of facial features

态变化,提取姿态指示的基础视觉特征;而同一旋转方向不同角度下,五官跨部位的联动约束规律,则为旋转位移关系表征,隐秘位移感知可提炼其内在几何约束特征。将三者融合,能全面构建头部姿态完整几何特征,既提升估计的精准性与抗遮挡能力,也为无约束场景下的头部姿态分析提供理论支撑。

2 本文模型

2.1 整体架构设计

基于五官旋转位移关系表征的头部姿态估计模型(ForHPE),核心架构由五官关系挖掘模块(Facial Relationship Mining Module, FRM)与隐秘位移特征表示模块(Covert Displacement Feature Representation Module,CDF)两部分构成(如图2所示),具体流程如下:首先,以2D人脸图像为输入,先通过特征提取器完成基础特征提取,将提取的特征图像进行分块操作后,展平为视觉Tokens随后将经过位置嵌入的视觉Tokens输入至FRM模块进行特征优化;特征增强后,进入CDF模块进行进一步的特征挖掘与损失优化,该模块在特征提取模块基础上扩展3个额外全连接层,结合细粒度姿态估计设计每个角度的分类损失与回归损失,并引入新的隐秘位移特征回归损失,该损失通过将遮挡图像的隐秘位移特征表示近似为非遮挡图像的特征表示,提升模型对遮挡场景的姿态预测能力;最后,通过损失计算与输出部分对各角度损失与隐秘位移特征损失进行加权求和,最终以欧拉角形式输出头部姿态预测结果。



图中虚线箭头表示具体展开部分,实线箭头表示数据流向。

图2 ForHPE模型的网络结构图

Fig. 2 Network architecture diagram of ForHPE model

2.2 五官关系挖掘模块

设x为原始特征图,使用BatchNorm对其进行归一化,得到$\hat{x}=\mathrm{BatchNorm}(x)$。

FRM包含了多尺度并行大卷积核模块(Multi Scale Parallel Large Convolution Kernel, MSPLCK)与增强并行注意力模块(Enhance Parallel Attention, EPA):前者可同步捕捉大面积遮挡区域并恢复特征细节,后者能并行提取原始特征中的共享全局信息与位置依赖型局部信息。

1) MSPLCK

MSPLCK采用三种不同卷积核尺寸(7×7、13×13、19×19)的并行扩张卷积可以提取多尺度特征。特征提取器因五官边缘信息突出而优先捕获其底层视觉特征,五官关系挖掘模块通过多尺度卷积(大卷积和中卷积具有类似于变换器中自注意力机制的长距离建模能力和大感受野,能够集中处理大面积的遮挡区域;小卷积则能够聚焦于小面积的遮挡区域并恢复纹理细节)、注意力机制及层级化融合,显式建模五官空间分布与跨部位依赖关系。将视觉Tokens进行层归一化后先后输入1×1卷积层和5×5卷积层,而后分别输入多尺度卷积层。将原始特征图的多尺度信息在通道维度上进

行拼接,使得x_3的特征维度是x的三倍,表示如下:

$$y = x + \mathrm{PWConv}(\mathrm{GELU}(\mathrm{PWConv}(F_m))) \quad (1)$$

其中,x表示原始输入特征图;F_m表示在通道维度上进行拼接后的多尺度信息。

将F_m输入到一个Pointwise卷积(PWConv)中,该感知机将F_m的特征维度转换为与x相同。这个PWConv包含两个逐点卷积,并使用高斯误差线性单元(Gaussian Error Linear Unit, GELU)作为激活函数。最后,PWConv的输出与恒等快捷连接x相加。PWConv不仅能够结合三种不同类型的特征,还能在拟合去遮挡特征中发挥作用。

2) EPA

EPA融合了不同类型的注意力机制,如图3所示,该模块包含简单的像素注意力(Simple Pixel Attention, SPA)、通道注意力(Channel Attention, CA)以及像素注意力(Pixel Attention, PA)。

SPA由特征提取分支和像素门控分支组成,两个分支的输出分别为PF_s和PA_s。将PF_s作为PA_s的像素门控信号,SPA模块的输出F_s计算方法为:

$$\begin{cases} PF_s = \mathrm{Conv}(\mathrm{PWConv}(\hat{x})) \\ PA_s = \mathrm{Sigmoid}(\mathrm{PWConv}(\hat{x})) \\ F_s = PF_s \otimes PA_s \end{cases} \quad (2)$$

其中,PWConv表示逐点卷积;Conv表示内核大小为3的卷积;Sigmoid函数用于提取全局像素门控特征。

PA包含一个PA_p分支,能够有效提取位置相关特征,例如图像中不同区域的遮挡。使用PA_p作为$\hat{x}$的全局像素门控信号,PA的输出F_p如(3)式所示:

$$\begin{cases} PA_p = \mathrm{Sigmoid}(f_1(\hat{x})) \\ F_p = \hat{x} \otimes PA_p \end{cases} \quad (3)$$

其中,$f_1(\hat{x})$表示使用PWConv-GELU-PWConv组合结构,该组合结构由两个逐点卷积层与中间GELU激活函数构成,用于拟合特征。

CA包含一个CA_c分支,能够高效地提取全局信息并改变特征的通道维度。使用CA_c对$\hat{x}$进行通道加权,得到全局通道门控信号(即CA的输出F_c)如(4)式所示:

$$\begin{cases} CA_c = \mathrm{Sigmoid}(f_2(\hat{x})) \\ F_c = \hat{x} \otimes CA_c \end{cases} \quad (4)$$

其中,$f_2(\hat{x})$代表使用PWConv-GELU-PWConv提取全局通道门控特征。

模型将3个不同的注意力门控结果在通道维度上进行拼接,得到EPA的输出F。通过多层感知机(Multi-Layer Perceptron, MLP),将拼接后的特征通道维度减少到与输入$\hat{x}$相同的维度。将MLP的输出(即PWConv(GELU(PWConv(F)))部分)与$\hat{x}$相加,最终得到FRM模块的输出y,如(5)式所示:

$$\begin{cases} F = \mathrm{Concat}(F_s, F_c, F_p) \\ y = \hat{x} + \mathrm{PWConv}(\mathrm{GELU}(\mathrm{PWConv}(F))) \end{cases} \quad (5)$$

CA能够更好地提取共享的全局信息并编码;PA则更擅长提取依赖位置的信息并编码。通过同时从原始特征中提取位置依赖信息和共享的全局信息,有望实现注意力机制的全局优化。然而当两种不同的注意力机制以串行方式使用时,无法达到全局最优:CA先对原始特征进行修改(提取全局信息),然后PA再对已修改的特征提取位置依赖信息。为了让增强注意力模块能够同时从原始特征中提取共享的全局变量和位置依赖的局部变量,EPA将3种不同的注意力模块并行化处理。具体来说,并行编码原始特征,将3条独立注意力通路的输出在通道维度上进行拼接,以获得组合特征F,并最终通过MLP对组合特征F进行融合得到y。该FRM模块能够更加有效地提取和整合遮挡情况下的特征信息,从而提升HPE的准确性和鲁棒性。

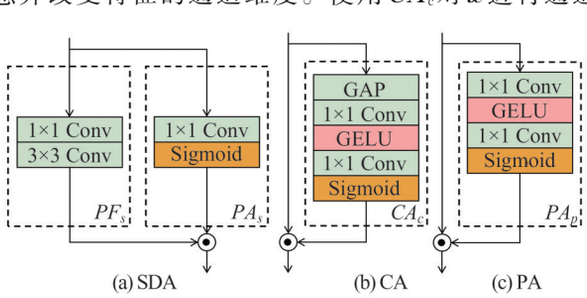


图3 简单像素注意力(SPA)、通道注意力(CA)和像素注意力(PA)的示意图

Fig. 3 Schematic diagrams of Simple Pixel Attention (SPA), Channel Attention (CA), and Pixel Attention (PA)

2.3 隐秘位移特征表示模块

在特征提取模块的基础上扩展出3个额外的全连接层,分别用于输出每个欧拉角的预测。骨干网络的输出会被展平为一个向量,成为后续各个全连接层的输入。每个全连接层的输出是一个logits向量,其中包含了针对预测角度的原始得分。在这个向量中,每个角度都被视为一个覆盖一定度数范围的区间。向量的大小取决于对特定欧拉角所划分的角度区间以及该欧拉角的完整预测范围。此外,这些全连接层的输出会应用到一个多损失策略中,以结合分类和回归两个部分,从而为给定的欧拉角生成最终的整体损失。对分类任务而言,在全连接层的K维输出向量上施加了Softmax激活函数和交叉熵损失。Softmax函数通过计算每个区间输出的指数e^{z_i},然后将其除以所有指数之和,将logits向量转换成概率$S(z_i)$,从而使得激活后的向量中所有概率的总和为1。$S(z_i)$如(6)所示:

$$S(z_i)=\frac{e^{z_i}}{\sum_{j=1}^{K}e^{z_j}} \tag{6}$$

其中,z_i是第i类的logit。

随后,交叉熵损失结果通过(7)式计算,得到分类损失L_{class}:

$$L_{\text{class}}=-\sum_{i=1}^{K}t_i\log(S(z_i)) \tag{7}$$

其中,t_i是地面真值(0或1)的得分激活结果。

除了分类损失之外,还引入了回归组件,用于确定并回归预测角度与真实值之间的误差(以度为单位)。通过使用Softmax激活获得的各区间的概率计算给定角度的期望,可以确定预测的角度$\hat{\theta}$:

$$\hat{\theta}=w\sum_{i=1}^{K}p_i\left(i-\frac{1+K}{2}\right) \tag{8}$$

其中,w是以度为单位的区间宽度;K是分类的区间数量;p_i是角度属于第i个区间的概率;偏移量$\frac{1+K}{2}$将区间索引移动到各自的区间中心。

用于回归组件的损失是预测角度$\hat{\theta}$与真实角度θ之间的均方误差,适用于N个预测,得到回归损失L_{reg}:

$$L_{\text{reg}}=\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}-\theta)^2 \tag{9}$$

分类损失L_{class}的目标是通过将每个姿态角度分类到一个角度区间中,帮助模型预测其附近的角度。随后,引入回归损失L_{reg}以辅助模型实现细粒度的角度预测。欧拉角的分类损失和回归损失通过一个正则化超参数α结合在一起,以控制这两项之间的平衡。总体而言,训练欧拉角时使用的3种损失可以表示为:

$$\begin{cases}L_{\text{yaw}}=L_{\text{class}}^{\text{yaw}}(\theta,\hat{\theta})+\alpha L_{\text{reg}}^{\text{yaw}}(\theta,\hat{\theta})\\ L_{\text{pitch}}=L_{\text{class}}^{\text{pitch}}(\theta,\hat{\theta})+\alpha L_{\text{reg}}^{\text{pitch}}(\theta,\hat{\theta})\\ L_{\text{roll}}=L_{\text{class}}^{\text{roll}}(\theta,\hat{\theta})+\alpha L_{\text{reg}}^{\text{roll}}(\theta,\hat{\theta})\end{cases} \tag{10}$$

其中,$\hat{\theta}$代表预测值;θ是训练数据集中提供的相应欧拉角真值,包含偏航角(Yaw)、俯仰角(Pitch)、翻滚角(Roll)。

2.4 损失函数

在头部姿态估计任务中,设计一个有效的损失函数是优化模型性能的关键。最大后验估计(Maximum A Posteriori, MAP)通过结合观测数据与先验分布推断最优的模型参数。为了提升头部姿态估计的精度与鲁棒性,本文基于MAP估计,结合KL散度和高斯先验设计了一种新的损失函数。给定一组人脸图像$\boldsymbol{X}$及其构建的真实角度分布$\hat{\boldsymbol{Y}}$,训练的目标是通过最大化后验概率$p(\beta|\boldsymbol{X},\hat{\boldsymbol{Y}})$找到最佳的$\beta$估计值。这里的$\beta$表示所提出的ForHPE模型中的所有神经网络参数,β^*表示所有神经网络参数的最佳估计值。β^*可以表示为:

$$\beta^*=\arg\max_{\beta}p(\beta|\boldsymbol{X},\hat{\boldsymbol{Y}}) \tag{11}$$

根据贝叶斯准则,(11)式可写作:

$$\beta^*=\arg\max\frac{p(\boldsymbol{X},\hat{\boldsymbol{Y}}|\boldsymbol{\beta})p(\beta)}{p(\boldsymbol{X},\hat{\boldsymbol{Y}})} \tag{12}$$

由于$p(\boldsymbol{X},\hat{\boldsymbol{Y}})$与$\beta$无关,因此可以认为$p(X,\hat{Y})$是一个常数。因此,(12)式可以写为:

$$\beta^*=\arg\max p(\boldsymbol{X},\hat{\boldsymbol{Y}}|\beta)p(\beta) \tag{13}$$

等式两边分别取对数函数,等式(13)可改写为:

$$\beta^*=\arg\max(\log p(\boldsymbol{X},\hat{\boldsymbol{Y}}|\beta)+\log p(\beta)) \tag{14}$$

需要定义两个概率密度函数。似然概率$p(\boldsymbol{X},\hat{\boldsymbol{Y}}|\beta)$表示预测分布与真实分布之间的距离。选择库尔贝克-莱布勒(KL)散度衡量这一距离。似然概率可以表示为:

$$p(\boldsymbol{X},\hat{\boldsymbol{Y}}|\beta)\propto\exp\left(-\sum_{t}y_t\ln\frac{y_t}{\hat{y}_t^*}\right) \tag{15}$$

其中,y_t为真实角度分布;$\hat{y}_t^*$表示模型预测的分布。在具体实现中,KL散度可近似为由角度损失三项组成的预测误差衡量项。为了抑制网络中噪声参数对结果的影响,对先验分布$p(\boldsymbol{\beta})$引入高斯正则化项:

$$p(\beta)=\frac{1}{\sqrt{2\pi}\,\sigma^2}\exp\left(-\frac{\beta^2}{2\sigma^2}\right) \tag{16}$$

将(15)式和(16)式代入(14)式,MAP问题转化为一个最小化对象的问题,该对象最小化概率的负对数。因此,(14)式可写作:

$$\beta^*=\arg\min\left(\sum_{t}y_t\ln\frac{y_t}{\hat{y}_t^*}+\omega\|\beta\|_2^2\right) \tag{17}$$

故提出的ForHPE模型的损失函数为:

$$L(\beta)=\sum_{t}y_t\ln\frac{y_t}{\hat{y}_t^*}+\omega\|\beta\|_2^2 \tag{18}$$

其中,符号ω表示正则化系数。在隐藏层中使用L2范数,以避免训练阶段参数的大幅增长。

此外,为了增强模型的泛化能力,尤其是在处理遮挡情况时,本文引入了隐秘位移特征回归损失。隐秘位移特征是包含神经网络最高层特征值的抽象低维空间。这些特征值编码了观察到的输入数据的最相关内部表示,并能够在隐秘位移特征中将语义上相似的姿态输出更靠近地放置。这种损失的背后理念是将遮挡图像的隐秘位移特征表示近似于非遮挡图像的隐秘位移特征表示,从而使模型能够更好地预测存在遮挡的姿态。额外的隐秘位移特征回归损失是遮挡图像的预测隐秘嵌入$\hat{E}$

与非遮挡图像的真实嵌入E_{gt}之间的均方误差。因此对于N个预测,表示如下:

$$L_{latent}=\frac{1}{N}\sum_{i=1}^{N}(\hat{E}-E_{gt})^2 \tag{19}$$

其中,非遮挡图像的真实隐秘位移特征嵌入E_{gt}是通过骨干网络上的推理之前提取的。最终总体损失是由角度损失和隐秘损失组成,并引入一个超参数γ,在此作为损失权重,来处理它们之间的权衡可以表示为:

$$L=(1-\gamma)(L_{yaw}+L_{pitch}+L_{roll})+\gamma L_{latent} \tag{20}$$

值得注意的是,本文的骨干网络有两个版本,分别为以ResNet50为骨干网络的ForHPE,和以Transformer为骨干网络的ForHPE+。基于五官旋转位移关系表征的高效头部姿态估计的模型训练过程如算法1所示。

算法1 基于五官旋转位移关系表征的高效头部姿态估计

输入: 输入图像$\boldsymbol{X}$,姿态标记θ_{gt},隐秘位移特征E_{gt},训练总轮数T

输出: 训练后ForHPE模型

#视觉Tokens构造

$\boldsymbol{V}=\{v_1,v_2,\cdots,v_n\}$

#位移Tokens构造

$\boldsymbol{D}=\{d_1,d_2,\cdots,d_n\}$

while 训练未达到设定的最大迭代次数 do:

  for j=1 to $(N+K)$ do:

    $\boldsymbol{F}$=Backbone($\boldsymbol{X}$)

    $\boldsymbol{F}_{FRM}$=FRM($\boldsymbol{V}$)

    $\hat{\boldsymbol{\theta}}$=CDF($\boldsymbol{D}$,$\boldsymbol{F}_{FRM}$)

    $L_{pose}=L_{class}+\alpha L_{reg}$

    $\hat{E}$=LatentPredict($\boldsymbol{F}$)

    L_{latent}=MSE($\hat{E}$,E_{gt})

    $L=(1-\gamma)L_{pose}+\gamma L_{latent}$

  end for

end while

return 训练后ForHPE模型

3 实验结果及讨论

3.1 数据集

本文的实验部分采用了3个开源的流行数据集(300W-LP、AFLW2000和BIWI数据集)。这些数据集包含精确头部姿态注释、众多身份不同的人脸以及不同的照明条件,并且都涉及大姿势的头部偏转图像,有利于在真实场景中对学习者的注意力进行评估。

300W-LP[34]:它是从300W数据集扩展而来的,其中包括68个关键点定位。300W-LP数据集的图像来自不同的数据集,包括AFW数据集、LFPW数据集、HELEN数据集、IBUG数据集和XM2VTS数据集等。共有61 225幅图像,其中1 786幅来自IBUG数据集,5 207幅来自AFW数据集,16 556幅来自LFPW数据集,37 676幅来自HELEN数据集。

AFLW2000[35]:它包含了使用面部建模技术3DMM拟合注释的AFLW数据集的前2 000幅图像。

BIWI[36]:该数据集包含24个由20人组成的序列,总共超过15 000幅图像。头部姿势的变化在偏航方向上为[−75°,+75°],在俯仰方向上为[−60°,+60°]。

合成遮挡数据集:采用Celestino等[11]工作中的多个合成遮挡数据集,这些数据集的生成过程依托于二维RGB-D图像数据。通过RGB-D相机同时捕获RGB信息和深度信息。

3.2 实验指标

本文选取评估头部姿态估计框架时广泛采用的指标偏航角(Yaw)误差、俯仰角(Pitch)误差、翻滚角(Roll)误差及平均绝对误差(Mean Absolute Error,MAE),作为评价指标。MAE表示如下:

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|\theta_i-\hat{\theta}_i\right|, \tag{21}$$

其中,$\hat{\theta}_i$是预测角度;θ_i为真实角度值。当比较不同方法的性能时,较小的Yaw角误差、Pitch角误差、Roll角误差、MAE值表示该方法具有更好的性能。

3.3 对比方法

为验证所提方法性能,选取HPE领域13种先进方法作对比,它们在损失设计、网络架构、特征处理等方面各有创新,覆盖不同技术路径以全面衡量效果。

HopeNet[30]:使用三个单独的损失对应三个欧拉角,每一个损失包含分类损失和回归损失两部分。

FSA-Net[33]:提出学习一种细粒度结构映射,在聚合前对特征进行空间分组。这种细粒度结构提供了基于部分的信息和池化值。

WHENet[31]:采用了HopeNet的整体框架。并采用更轻的主干网络,结合了来自MobileNetV2的倒排残差块,以减少参数数量,同时增加跳过连接。

HeadPosr[21]:首次将Transformer用于HPE领域,探究Transformer编码器的作用。

TirNet[15]:提出利用旋转矩阵中的三个向量作为头部姿态估计的表示。

DAD-3D[12]:提出了3D密集头部校准野外模型,利用丰富的监督和多个相关任务的协同训练。

LwPosr[18]:采用深度可分离卷积和Transformer编码器层的混合结构,该结构分为两个流和三个阶段,以提供精细的回归预测头部姿态。

TokenHPE[20]:出了一种基于Transformer架构的新颖的关键少数关系感知方法,该方法可以学习面部部分之间的关系。

6DRepNet[16]:通过引入旋转矩阵形式来处理旋转标签的模糊问题,并提出了一种连续的六维旋转矩阵表示,以实现高效且稳健的直接回归。

LSR[11]:提出了一种新颖的深度学习方法,用于解决遮挡下的头部姿态估计问题。

CMSA-L[22]:提出了一种新的注意力机制,称为级联多尺度注意力(CMSA),专为CNN-ViT混合架构设计,以有效处理低分辨率输入。

ORC[13]:通过引入面部landmark加权图的Ricci曲率离散化形式来处理头部姿态估计中对光照、遮挡等环境变化敏感且传统特征难以捕捉全局几何关联的问题,并提出了一种基于Ollivier-Ricci和Forman-Ricci两种离散曲率的特征提取与随机森林回归结合的框架,以实现高效且稳健的单图像头部姿态估计。

TUL[14]:提出了一个新的框架,它包括一个用于联合真实标签和姿态预测的多任务网络、以及一组可学习的三维标准地标和一个图像生成网络。通过对条件图像生成和几何一致性的综合损失,对未标记的人脸图像进行协同学习。

3.4 实验设置

ForHPE模型实验中,输入图像尺寸设为224×224,分别加载合成遮挡与原始300W-LP数据集图像;批大小均为64,ViT版本训练100个训练轮数,ResNet50版本训练25个训练轮数。采用Adam优化器,初始学习率1E-5,超参数α=2、γ=0.999。训练用300W-LP数据集遮挡版本,测试用BIWI和AFLW2000数据集的遮挡与非遮挡版本;在不借助跟踪手段的前提下,针对旋转角度在[-99°,99°]范围内的人脸,采用MTCNN算法[37]完成人脸检测。实验在Linux操作系统下进行,使用NVIDIA A800 GPU,深度学习框架为PyTorch。

3.5 实验结果与分析

3.5.1 对比实验

为对比模型性能,进行不同模型在AFLW2000、BIWI数据集上性能测试的对比实验。在AFLW2000数据集上测试的实验结果如表1所示。本研究提出的基于ResNet的ForHPE在遮挡处理后的AFLW2000数据集上,偏航角误差仅4.049°,是该场景下所有对比模型中这一指标的最佳结果,仅在遮挡场景的Pitch、Roll指标上略逊于LSR模型,本研究提出的基于Transformer的ForHPE+模型在该场景下的俯仰角、翻滚角误差及综合MAE较优。在无遮挡的AFLW2000数据集上,ForHPE的Yaw角度误差3.346°是该场景下的最低值,ForHPE+的Yaw误差3.553°也仅略高于这一最优结果,且其Pitch误差5.125°也优于不少对比方法。

在300W-LP数据集上训练,在BIWI数据集上

表1 遮挡和非遮挡AFLW2000数据集上与众多模型的比较结果

Table 1 Comparison results with many models on occluded and non occluded AFLW2000 datasets

(°)

模型	主干网络	遮挡处理后的AFLW2000数据集				无遮挡的AFLW2000数据集			
		Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
HopeNet[30]	ResNet50	12.438	10.277	8.586	10.434	4.965	5.250	3.956	4.724
FSA-Net[33]	CNN	13.664	10.880	10.067	11.537	5.109	6.462	3.356	5.642
WHENet[31]	CNN	16.515	11.669	10.267	12.817	4.475	6.222	5.017	5.238
HeadPosr[21]	Transformer	—	—	—	—	4.640	5.840	4.300	4.920
TirNet[15]	ResNet50	—	—	—	—	4.040	5.780	4.200	4.670
DAD-3D[12]	ResNet50	7.084	14.953	13.680	11.906	3.667	**4.765**	**3.153**	**3.861**
LwPosr[18]	Transformer	10.771	9.278	8.335	9.461	4.267	5.015	3.722	4.335
TokenHPE [20]	Transformer	—	—	—	—	4.530	5.730	4.290	4.850
6DRepNet[16]	ResNet50	8.904	9.799	7.408	8.704	3.630	<u>4.918</u>	<u>3.371</u>	<u>3.979</u>
LSR[11]	ResNet50	4.902	<u>5.968</u>	<u>4.625</u>	5.165	3.754	5.210	3.819	4.261
CMSA-L[22]	CNN-ViT	—	—	—	—	4.150	5.100	3.890	4.380
ORC[13]	ResNet50	—	—	—	—	3.350	4.390	3.650	3.790
TUL[14]	CNN	—	—	—	—	3.460	4.500	3.700	3.890
ForHPE	ResNet50	**4.049**	6.437	4.926	<u>5.137</u>	**3.346**	5.768	4.185	4.433
ForHPE+	Transformer	<u>4.441</u>	**5.822**	**4.567**	**4.943**	<u>3.553</u>	5.125	3.798	4.159

注:“—”表示该模型未在此数据集上进行实验;加粗数据表示该指标最优结果;下划线数据表示该指标次优结果。

测试的实验结果如表2所示,与AFLW2000数据集中对比的方法相同。由实验结果可直观发现,本文提出的ForHPE+在存在面部遮挡的BIWI数据集上表现最优,其Roll、Pitch、Yaw三个姿态角的平均绝对误差(MAE)均为所有对比方法中最低。对于基础模型ForHPE,其整体性能弱于改进版ForHPE+;但与其余主流方法横向对比后可知,仅LSR模型整体误差略低于ForHPE,其余对比方法的姿态估计误差均高于ForHPE。在无遮挡的BIWI数据集上,ForHPE+虽未取得全局最优结果,但仍具备极具竞争力的较优性能;其中该模型Pitch维度的角度误差在全部对比算法中达到最低。同时ForHPE+模型的总体MAE误差在所有对比方法中是第二低的,与效果最好的方法只相差0.234°。

综合实验结果表明,ForHPE+模型在遮挡场景下表现优于现有先进方法,在无遮挡场景中也保持了较强竞争力,充分验证了五官旋转位移表征与双模块设计的有效性。

表2 遮挡和非遮挡BIWI数据集上与众多模型的比较结果

Table 2 Comparison results with many models on occluded and non occluded BIWI datasets

(°)

模型	主干网络	遮挡处理后的BIWI数据集				无遮挡的BIWI数据集			
		Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
HopeNet[30]	ResNet50	6.725	8.616	7.338	7.560	4.375	3.559	3.348	3.761
FSA-Net[33]	CNN	10.987	9.848	7.846	9.560	5.420	5.568	4.515	5.168
WHENet[31]	CNN	21.101	13.188	11.225	15.171	4.395	4.346	4.674	4.472
HeadPosr[21]	Transformer	—	—	—	—	4.080	5.100	3.020	4.060
TirNet[15]	ResNet50	—	—	—	—	4.110	4.760	3.050	3.970
DAD-3D[12]	ResNet50	5.532	7.924	7.478	6.978	3.981	5.241	2.923	4.048
LwPosr[18]	Transformer	10.784	9.654	10.309	10.249	4.417	4.875	3.191	4.161
TokenHPE [20]	Transformer	—	—	—	—	3.950	4.510	2.710	3.720
6DRepNet[16]	ResNet50	7.883	14.983	9.665	10.844	**3.241**	4.482	**2.683**	**3.472**
LSR[11]	ResNet50	5.762	5.529	3.997	5.096	4.366	3.686	3.786	3.946
CMSA-L[22]	CNN-ViT	—	—	—	—	4.490	4.630	2.850	3.990
ORC[13]	ResNet50	—	—	—	—	3.570	4.190	3.850	3.870
TUL[14]	CNN	—	—	—	—	3.600	8.500	8.700	6.900
ForHPE	ResNet50	5.726	7.043	5.008	5.926	4.023	4.768	3.759	4.183
ForHPE+	Transformer	**5.361**	**4.813**	**3.634**	**4.603**	4.112	**3.533**	3.473	3.706

注:“—”表示该模型未在此数据集上进行实验,表中加粗数据表示该指标最优结果,下划线数据表示该指标次优结果。

3.5.2 消融实验

1) 损失权重γ

为进一步探讨不同超参数值γ中哪一个的效果最佳,进行了损失权重γ的消融实验。在ForHPE+模型上进行实验,在300W-LP数据集上进行训练,在遮挡和未遮挡的AFLW2000数据集和BIWI数据集上对比MAE以及Yaw、Pitch、Roll三个角度的误差,找出最佳的超参数γ的值。共讨论了5种超参数γ值的设置,分别为{0, 0.5, 0.990, 0.999, 1}。通过手动调试可发现明显变化趋势,该超参数在数值接近1的区间内模型性能达到最优,因此本次超参数的选取策略以实验观察得到的规律为依据。实验结果如图4所示,当γ取值为0.999时,模型在所有测试数据集上的MAE及各角度误差大多达到最小值,是兼顾角度预测精度与抗遮挡鲁棒性的最优超参数设置。对比不同超参数γ的三维柱状图,γ=0.999时在遮挡和未遮挡的AFLW2000、BIWI数据集上均取得了最优或接近最优的误差指标。这说明在CDF中,给隐秘位移特征回归损失赋予较高的权重最有利于整体性能。通过对损失权重γ的消融实验可以发现,当γ取接近1的最优值时,即0.999,模型在各类测试场景下均能达到最佳精度与鲁棒性,验证了对隐秘位移特征赋予较高权重的合理性与必要性。

此时模型既能充分利用隐秘位移特征表示对齐来纠正遮挡对特征学习造成的干扰,又不会因完全忽略角度损失而影响基本的姿态预测精度,从而达到遮挡场景和正常场景的较好平衡。

2) 关键模块的有效性

模型主要由FRM和CDF两个部分组成,为了探讨FRM、CDF在模型中的有效性,在ForHPE+模型上进行实验,在300W-LP数据集上进行训练,在遮挡

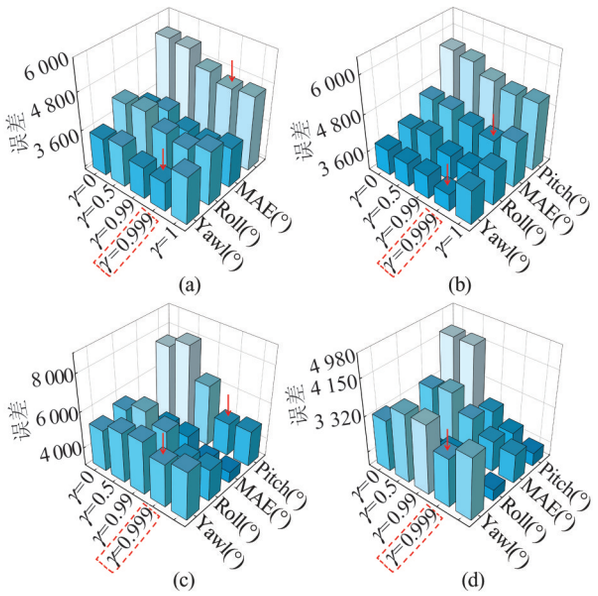


图4 对不同超参数γ的三维柱状图(a)遮挡处理后的AFLW2000数据集;(b)无遮挡的AFLW2000数据集;(c)遮挡处理后的BIWI数据集;(d)无遮挡的BIWI数据集

Fig. 4 Three-dimensional histogram of different hyper parameter γ (a) AFLW2000 dataset after occlusion processing; (b) An unobstructed AFLW2000 dataset; (c) BIWI dataset after occlusion processing; (d) Unobstructed BIWI dataset

处理后的AFLW2000数据集上测试。对比MAE以及Yaw、Pitch、Roll三个角度的误差,验证关键模块的有效性。FRM模块中包含MSPLCK和EPA两个重要组成部分,具体讨论FRM的两个子模块,如表3所示。

可以看出在基线模型上逐步加入各个关键模块后,模型的Yaw、Pitch、Roll以及MAE均有不同程度的下降,说明每个模块都对提升模型性能发挥了积极作用。其中,当MSPLCK、EPA、CDF同时启用时,指标均达到最低,表现最优。这样的性能提升主要得益于MSPLCK在特征层面同时捕捉不同尺度的遮挡信息,EPA让模型并行关注全局通道特征和局部空间信息。这充分验证了FRM模块作为模型核心特征提取单元的设计合理性,其内部子模块的协同作用实现了对遮挡场景的针对性适配,强化了五官关系的细粒度建模。

表3 在遮挡处理后的AFLW2000上的消融实验

Table 3 Ablation Experiment of key modules on AFLW2000 after occlusion processing

FRM		CDF	Yaw/(°)	Pitch/(°)	Roll/(°)	MAE/(°)	Δ/%
MSPLCK	EPA						
×	×	×	5.427	6.331	4.618	5.459	−10.4
×	×	√	5.138	6.132	4.701	5.324	−7.7
×	√	×	5.104	6.017	4.659	5.256	−6.3
×	√	√	5.013	5.932	4.597	5.181	−4.8
√	√	√	**4.441**	**5.822**	**4.567**	**4.943**	—

注:"Δ"表示相对于MSPLCK、EPA、CDF均启用的完整模型MAE的变化百分比,负值表示劣于完整模型;加粗数据表示该指标最优结果;下划线数据表示该指标次优结果。

3) 300W-LP数据集最佳遮挡比

实验是在合成遮挡数据集300W-LP数据集上进行实验,故训练时修改遮挡图像和原始图像的批大小调整遮挡比。为了探讨不同遮挡比的设置对实验结果的影响,进行了最佳遮挡比实验。本小节共讨论了3种遮挡比的设置,分别为遮挡图像与原始图像的比例为1:1、遮挡图像与原始图像的比例为1:2、使用原始的300W-LP数据集。通过实验验证不同遮挡比对隐秘位移特征回归的影响,同样在表现较好的ForHPE模型上进行实验,在300W-LP数据集上进行训练,在遮挡和未遮挡的AFLW2000数据集和BIWI数据集上对比MAE以及三个角度的误差,来论证最佳遮挡比。结果如表4和表5所示,可以清晰观察到不同遮挡程度下模型的性能变化规律,其中遮挡图像与原始图像比例为1:1的中

表4 不同遮挡比在AFLW2000的实验结果对比

Table 4 Experimental results of different occlusion ratios in occluded and unobstructed AFLW2000

(°)

遮挡比	Yaw	Pitch	Roll	MAE
	遮挡处理后的AFLW2000数据集			
1:1	**4.441**	**5.822**	**4.567**	**4.943**
1:2	4.869	6.079	4.670	5.206
原始数据集	4.835	5.846	4.469	5.050
	无遮挡的AFLW2000数据集			
1:1	**3.553**	**5.125**	**3.798**	**4.159**
1:2	3.789	5.295	3.908	4.331
原始数据集	3.863	5.125	3.826	4.271

注:加粗数据表示该指标最优结果;下划线数据表示该指标次优结果。

表5 不同遮挡比在BIWI的实验结果对比

Table 5 Experimental results of different occlusion ratios in occluded and unobstructed BIWI

(°)

遮挡比	Yaw	Pitch	Roll	MAE
	遮挡处理后的BIWI数据集			
1:1	**5.361**	**4.813**	**3.634**	**4.603**
1:2	6.425	5.621	3.972	5.339
原始数据集	6.052	4.953	4.371	5.125
	无遮挡的BIWI数据集			
1:1	**4.112**	3.533	3.473	**3.706**
1:2	4.630	3.661	**3.346**	3.879
原始数据集	4.462	**3.190**	3.708	3.787

注:加粗数据表示该指标最优结果;下划线数据表示该指标次优结果。

度遮挡场景,成为模型性能表现的最优区间。在该比例下,无论是遮挡测试集还是未遮挡测试集,模型在AFLW2000与BIWI两大基准数据集上均取得最佳或接近最佳的误差表现,充分体现了模型对中度遮挡场景的良好适配性。

当遮挡比达到1:1,进入中度遮挡场景时,尽管部分面部关键特征受到遮蔽,带来一定的特征提取困难,但模型依靠训练数据中对五官旋转位移关系的建模,充分挖掘剩余有效面部信息,实验指标提升。这种性能变化趋势直观反映了模型在不同遮挡强度下的鲁棒性差异,也印证了1:1的中度遮挡训练比例能让模型在特征学习与抗遮挡能力之间达到最佳平衡。

3.5.3 可视化效果与分析

1) t-SNE隐秘位移特征可视化

隐秘位移特征是一个抽象的多维空间,用于映射神经网络对外部输入数据的相关内部表征,并提供对训练所学内容的洞察。在这个空间中,数据点之间的距离与输入数据的语义相似度相关。因此,相似的样本在该空间中彼此距离更近,而不相似的样本距离更远。为了直观观察隐秘位移特征,进行了t-SNE隐秘位移特征可视化实验。使用的模型ForHPE和ForHPE+在无遮挡的AFLW2000数据集上进行了推理,并保存了输出的隐秘位移特征,以应用t-分布随机邻居嵌入(t-SNE)技术。它是一种非线性方法,能够通过为每个数据点在二维或三维空间中指定一个位置可视化高维数据如图5所示。

图5(a)和图5(b)颜色相近(表示头部姿态角度相近)的点在二维空间中更为接近,形成较明显的簇。这说明模型在其隐秘位移特征中,已经能够将姿态相似的样本映射到相邻区域。颜色差异较大的样本(姿态差距较大)分布在二维平面的不同位置,彼此间距离更远,表明模型在区分各类极端或差异较大的头部角度时具有一定的判别能力。同时观察左、右两个子图可以发现,尽管两种模型在可视化分布上有些差异,但都呈现出随角度变化由中心向外或在不同区域分散的趋势。这充分验证了所提模型学习到的隐秘位移特征具备良好的姿态聚类与角度区分特性,能够有效表征头部姿态的细微变化与全局差异。

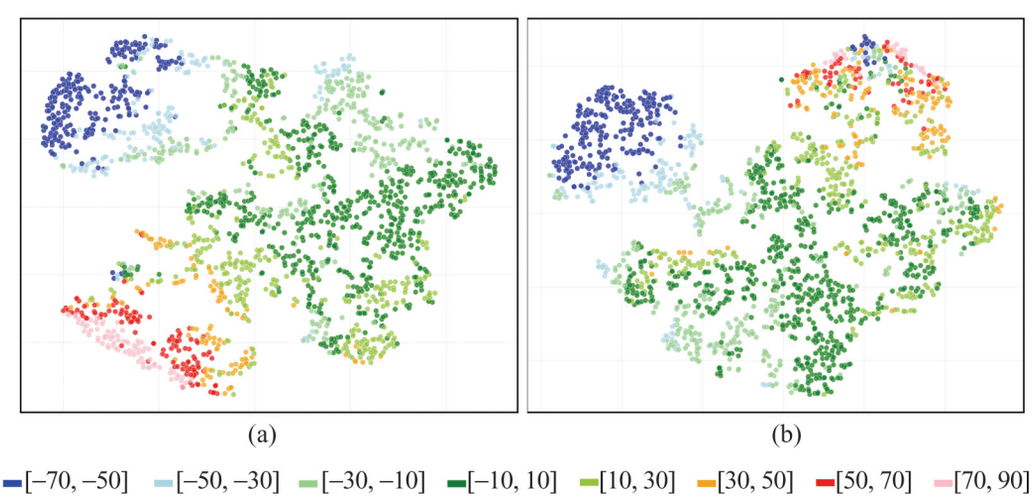


图5 无遮挡的AFLW2000数据集上两种模型的可视化(a)ForHPE;(b)ForHPE+

Fig. 5 Visualization of two models on non occluded AFLW2000 dataset(a)ForHPE;(b)ForHPE+

2) 深度特征可视化

为了观察对比不同模型提取特征位置,进行了深度特征可视化实验。为避免可视化图像排布繁杂、对比效果模糊,本实验从13种对比方法中挑选3种综合性能较优的代表性算法开展对照;图中虚线圈内区域代表模型重点关注的区域,可直观区分各模型的特征聚焦差异。图6展示了在AFLW2000数据集上,各个模型的深度特征可视化。

可以观察到,相较于6DrepNet、LwPosr和TokenHPE,ForHPE模型能够更准确地聚焦在人脸关键区域,在面对侧脸、低头及遮挡等复杂场景时仍表现出较强的关键区域建模能力。这表明ForHPE在引入FRM模块后,显著提升了模型对面部关键结构的感知能力,能够更精细地捕捉面部边缘轮廓与器官分布信息,从而增强了姿态估计的准确性与在复杂环境下的鲁棒性。这充分印证了FRM模块在强化面部关键特征感知、抑制无关干扰方面的核心作用,其设计思路与头部姿态估计任务的核心需求契合,为模型在复杂场景下保持优异性能提供了特征支撑。

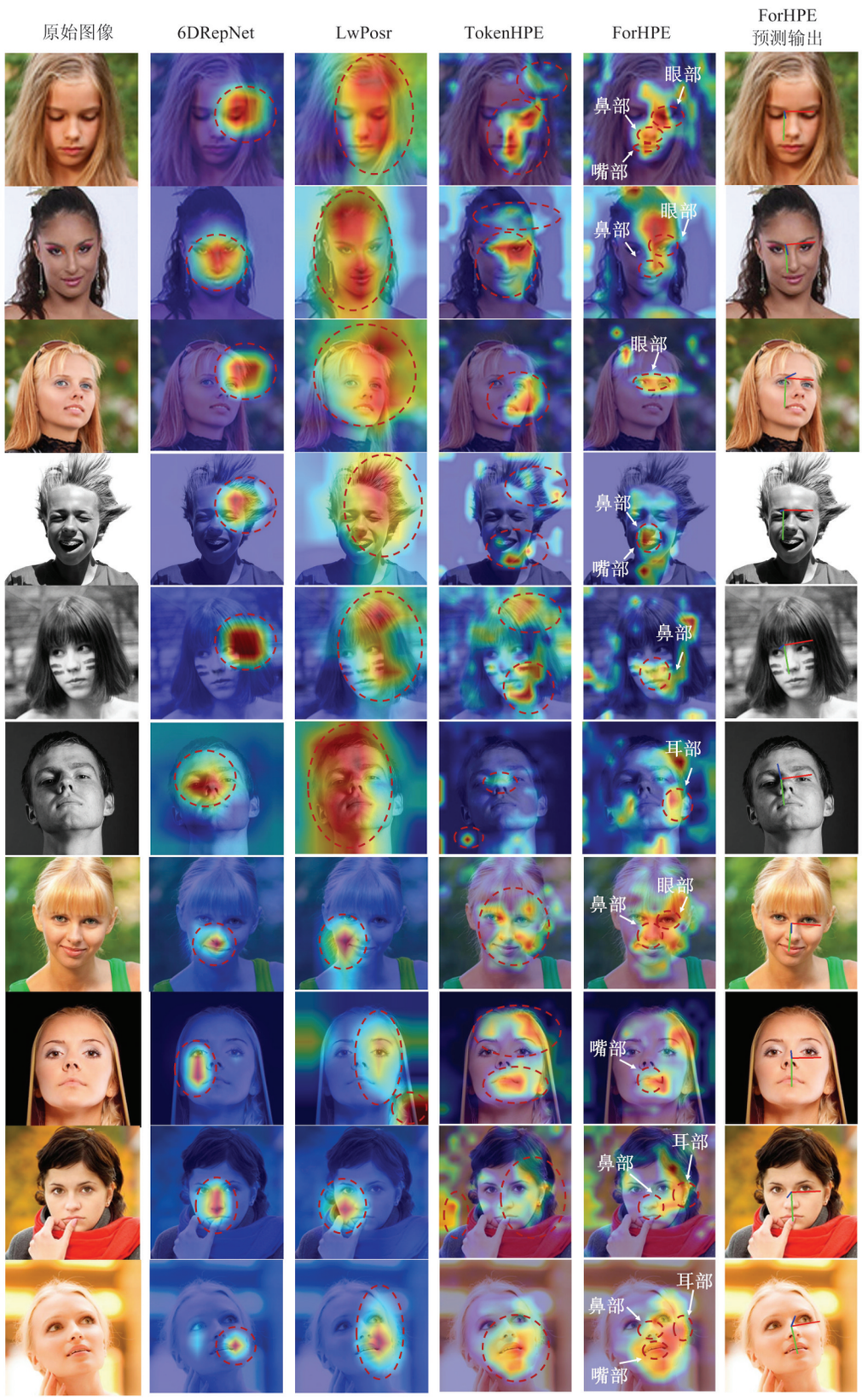


图6 AFLW2000数据集上,各模型深度特征可视化

Fig. 6 Visualization of depth features of each model on AFLW2000 dataset

3)注意力图像迭代可视化

为了探讨所提HPE模型在训练过程中的关键关注区域,本文开展了注意力图像迭代可视化实验。图7从左向右展示了在AFLW2000数据集上,ForHPE模型通过注意力图像可视化的训练过程,图中虚线圈内区域代表模型分配高权重、重点捕捉

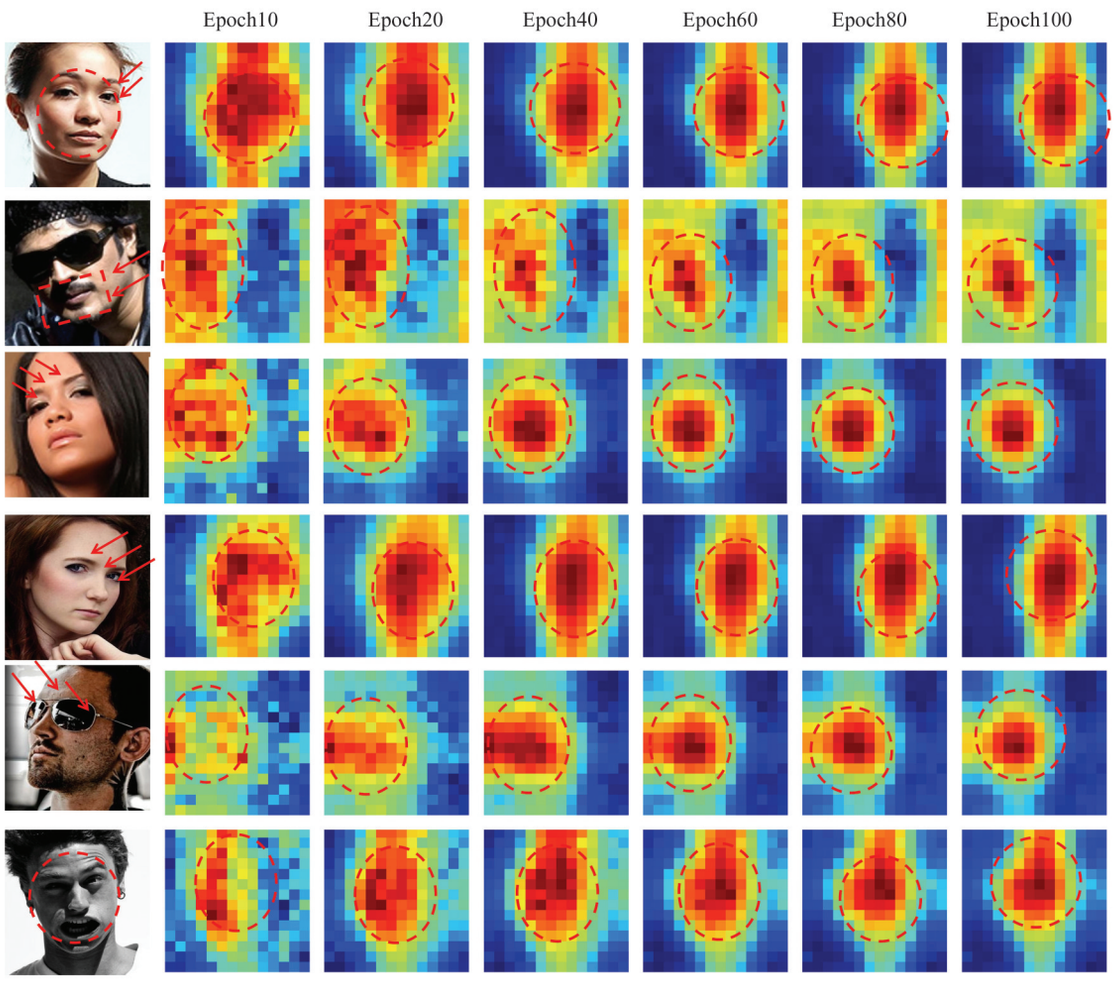


图7 AFLW2000数据集上,各模型深度特征可视化

Fig. 7 Visualization of depth features of each model on AFLW2000 dataset

的面部区域,可直观观察模型特征聚焦的变化规律。由图7可见,模型在迭代过程中在面部特征和头部轮廓等关键区域展示出较高的注意力。这是因为这些区域对于头部姿态的判定起到了重要作用,尤其是在姿态角度较为极端的情况下。

针对存在大面积面部遮挡、头部极端偏转角度的困难样本,模型能够自适应调整注意力分布策略,主动降低被遮挡无效区域的权重占比,将注意力资源集中分配至剩余未被遮挡、富含姿态判别信息的有效面部区域,以此规避遮挡带来的特征干扰。该差异化优势得益于FRM五官关系挖掘模块与CDF隐秘位移表征模块的协同配合,二者联合增强了模型筛选、锁定高价值面部关键区域的感知能力,能够过滤冗余干扰信息,稳定提取可靠的姿态特征,最终显著提升模型在遮挡、极端偏转等复杂无约束场景下的头部姿态估计精度与鲁棒性。

4) 挑战情景下的可视化

为了验证本文模型在挑战性场景中的头部姿态估计表现,本节对AFLW2000包含复杂场景的测试图像进行了可视化分析。如图8所示,本节对比各种方法在挑战情景下的模型能力,分别对比模型预测结果和热力图,热力图内虚线圈内区域对应模型核心关注区域,便于横向对比各模型的注意力分布特点。

在低光照、耳部遮挡、头部极端偏转等易造成面部有效信息缺失、成像质量受损的挑战性场景下,本文ForHPE模型,依托五官旋转位移关联特征建模的优势,热力图虚线圈内区域能够高度贴合眼、鼻、耳、口等具备姿态几何信息的面部关键结构,热区边界清晰、聚焦范围紧凑,几乎不存在向无关背景扩散的无效注意力。

多组挑战情形下样本的可视化结果直观证实,本文所提模型在头部姿态估计任务中具备更强的场景鲁棒性。面对光照不足、局部遮挡、极端偏转等复杂干扰条件时,模型可稳定锁定与头部姿态强相关的五官特征区域,有效规避无效背景信息带来的特征干扰,为各类无约束复杂场景下精准的头部姿态预测提供稳健、高辨识度的特征表征支撑。

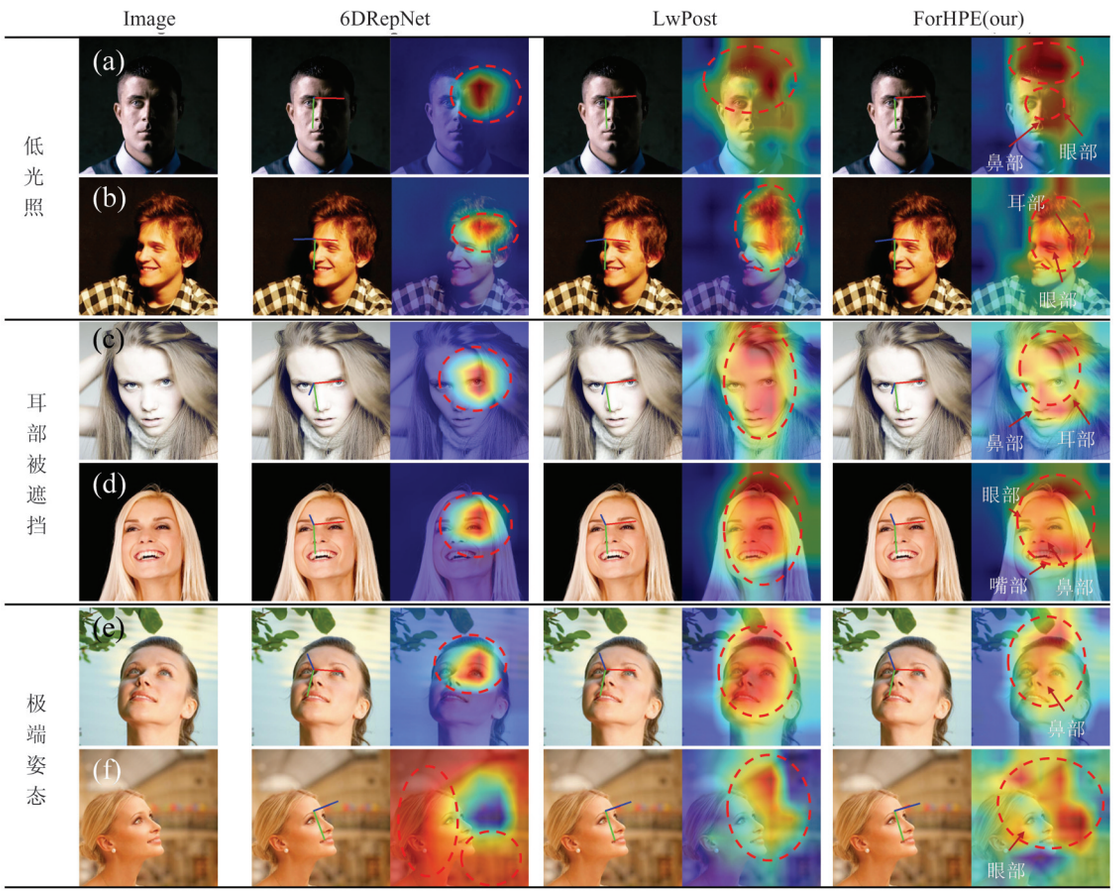


图8 挑战情景下的可视化

Fig. 8 Visualization in challenge scenarios

4 结 语

本文提出了一种基于五官旋转位移关系表征的HPE方法,旨在充分挖掘五官之间的空间结构与高层语义关联,从而增强模型对遮挡情况下关键区域关系的建模能力。本文设计了FRM模块,利用五官结构可见性与准确性之间的负相关特性,引导模型关注五官间的空间几何与语义关系。提出了CDF模块,以旋转过程中五官间的相对位移为核心构建稳健的特征表示,适应复杂遮挡场景,并引入隐秘位移特征回归组件,实现对遮挡图像特征向非遮挡图像特征的迁移。最后,在BIWI与AFLW2000两个数据集及其遮挡版本上进行实验,结果表明本文方法在遮挡条件下显著优于多种先进方法,在非遮挡场景中也保持了竞争性的准确率,验证了所提模型的有效性。后续将基于本文头部姿态估计模型开展多方向拓展研究,充分挖掘模型提取面部空间关联特征的优势,将现有框架迁移至人脸年龄识别、面部表情识别等相关人脸分析任务中,进一步丰富模型的应用场景与研究价值。

参考文献:

[1] TOURANI S, TOURANI S, MAHMOOD A, *et al*. Unsupervised discovery of facial landmarks and head pose [C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2025: 21192-21202. DOI: 10.1109/CVPR52734.2025.01974.

[2] BOSCH N, D' MELLO S K. Automatic detection of mind wandering from video in the lab and in the classroom [J]. *IEEE Transactions on Affective Computing*, 2021, **12**(4): 974-988. DOI:10.1109/TAFFC.2019.2908837.

[3] JIANG J X, STRELI P, MEIER M, *et al*. EgoPoser: robust real-time egocentric pose estimation from Sparse and Intermittent observations everywhere [C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 277-294. DOI:10.1007/978-3-031-72627-9_16.

[4] ROTH M, GAVRILA D M. intrApose: Monocular driver 6 DOF head pose estimation leveraging camera intrinsics[J]. *IEEE Transactions on Intelligent Vehicles*, 2023, **8**(8): 4057-4068. DOI:10.1109/TIV.2023.3274068.

[5] CHOTIKKAKAMTHORN K, LIE W N, RITTHIPRAVAT P, *et al*. Medical application of deep-learning-

based head pose estimation from RGB image sequence[J]. *Computers in Biology and Medicine*, 2025, **195**: 110620. DOI:10.1016/j.compbiomed.2025.110620.

[6] 章毅, 吕嘉仪, 兰星, 等. 结合面部动作单元感知的三维人脸重建算法[J]. 软件学报, 2024, **35**(5): 2176-2191. DOI:10.13328/j.cnki.jos.007029.
ZHANG Y, LÜ J Y, LAN X, *et al*. AU-aware algorithm for 3D facial reconstruction [J]. *Journal of Software*, 2024, **35**(5): 2176-2191. DOI: 10.13328/j.cnki.jos.007029(Ch).

[7] HEMPEL T, ABDELRAHMAN A A, AL-HAMADI A. Toward robust and unconstrained full range of rotation head pose estimation [J]. *IEEE Transactions on Image Processing*, 2024, **33**: 2377-2387. DOI: 10.1109/TIP.2024.3378180.

[8] BARRA P, BARRA S, BISOGNI C, *et al*. Web-shaped model for head pose estimation: An approach for best exemplar selection[J]. *IEEE Transactions on Image Processing*, 2020, **29**: 5457-5468. DOI: 10.1109/TIP. 2020.2984373.

[9] 李俊, 袁通达, 陈黎. 基于多尺度与多级语义融合Transformer的人体姿态估计[J]. 武汉大学学报(理学版), 2025, **71**(4): 473-484. DOI: 10.14188/j. 1671-8836.2024.0094.
LI J, YUAN T D, CHEN L. Human pose estimation based on multi-scale and multi-level semantic fusion transformer[J]. *Journal of Wuhan University* (*Natural Science Edition*), 2025, **71**(4): 473-484. DOI:10.14188/j.1671-8836.2024.0094(Ch).

[10] YU Y, MORA K A F, ODOBEZ J M. HeadFusion: 360° head pose tracking combining 3D morphable model and 3D reconstruction[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(11): 2653-2667. DOI:10.1109/TPAMI.2018.2841403.

[11] CELESTINO J, MARQUES M, NASCIMENTO J C, *et al*. 2D Image head pose estimation *via* latent space regression under occlusion settings[J]. *Pattern Recognition*, 2023, **137**: 109288. DOI:10.1016/j.patcog.2022.109288.

[12] MARTYNIUK T, KUPYN O, KURLYAK Y, *et al*. DAD-3DHeads: A large-scale dense, accurate and diverse dataset for 3D head alignment from a single image [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2022: 20910-20920. DOI: 10.1109/CVPR52688.2022.02027.

[13] ABATE A F, CASCONE L, NAPPI M. Ricci curvature discretizations for head pose estimation from a single image [J]. *Pattern Recognition*, 2025, **166**: 111648. DOI: 10.1016/j.patcog.2025.111648.

[14] ZOU Z M, JIA D, TANG W. Towards unsupervised learning of joint facial landmark detection and head pose estimation [J]. *Pattern Recognition*, 2025, **162**: 111393. DOI:10.1016/j.patcog.2025.111393.

[15] CAO Z W, CHU Z C, LIU D F, *et al*. A vector-based representation to enhance head pose estimation[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). New York: IEEE Press, 2021: 1187-1196. DOI:10.1109/WACV48630.2021.00123.

[16] HEMPEL T, ABDELRAHMAN A A, AL-HAMADI A. 6d rotation representation for unconstrained head pose estimation [C]//2022 IEEE International Conference on Image Processing (ICIP). New York: IEEE Press, 2022: 2496-2500. DOI: 10.1109/ICIP46576.2022.9897219.

[17] SHEKA A, SAMUN V. Boosting of head pose estimation by knowledge distillation [EB/OL]. [2022-02-28]. https://arxiv.org/pdf/2108.09183.

[18] DHINGRA N. LwPosr: Lightweight efficient fine grained head pose estimation[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). New York: IEEE Press, 2022: 1204-1214. DOI: 10.1109/WACV51458.2022.00127.

[19] LIU H, LIU T T, ZHANG Z L, *et al*. ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human - computer interaction [J]. *IEEE Transactions on Industrial Informatics*, 2022, **18**(10): 7107-7117. DOI:10.1109/tii.2022.3143605.

[20] ZHANG C, LIU H, DENG Y J, *et al*. TokenHPE: Learning orientation tokens for efficient head pose estimation *via* transformers [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2023: 8897-8906. DOI: 10.1109/CVPR52729.2023.00859.

[21] DHINGRA N. HeadPosr: End-to-end trainable head pose estimation using transformer encoders [C]//2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). New York: IEEE Press, 2021: 1-8. DOI: 10.1109/FG52635.2021.9667080.

[22] LU X Y, SUGANUMA M, OKATANI T. Cascaded multi-scale attention for enhanced multi-scale feature extraction and interaction with low-resolution images [PP/OL]. *V4. arXiv* (2026-07-13). https://doi. org/10.48550/arXiv.2412.02197.

[23] 李成龙, 钟凡, 马昕, 等. 基于卡尔曼滤波和随机回归森林的实时头部姿态估计[J]. 计算机辅助设计与图形学学报, 2017, **29**(12): 2309-2316. DOI:10.3724/SP.J.1089.2017.16521.

LI C L, ZHONG F, MA X, *et al*. Real-time head pose estimation based on Kalman filter and random regression forest[J]. *Journal of Computer—Aided Design & Computer Graphics*, 2017, **29**(12): 2309-2316. DOI: 10.3724/SP.J.1089.2017.16521(Ch).

[24] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010. DOI:10.5555/3295222.3295349.

[25] 刘婷婷, 王敏红, 邓李茜, 等. 基于褶纹基元关系感知的高效面部表情识别方法[J]. 武汉大学学报(理学版), 2025, **71**(5): 652-666. DOI: 10.14188/j. 1671-8836. 2025.0014.

LIU T T, WANG M H, DENG L Q, *et al*. Pleated primitive-aware semantic relationship learning for efficient facial expression recognition *via* transformers[J]. *Journal of Wuhan University* (*Natural Science Edition*), 2025, **71**(5): 652-666. DOI: 10.14188/j.1671-8836.2025.0014 (Ch).

[26] ESCOBAR M, PUENTES J, FORIGUA C, *et al*. EgoCast: Forecasting egocentric human pose in the wild[C]//2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). New York: IEEE Press, 2025: 5831-5841. DOI: 10.1109/WACV61041.2025.00569.

[27] KHAN K, MAURO M, MIGLIORATI P, *et al*. Head pose estimation through multi-class face segmentation [C]//2017 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE Press, 2017: 175-180. DOI:10.1109/ICME.2017.8019521.

[28] GUPTA A, THAKKAR K, GANDHI V, *et al*. Nose, eyes and ears: Head pose estimation by locating facial keypoints [C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE Press, 2019: 1977-1981. DOI:10.1109/ICASSP.2019.8683503.

[29] VALLE R, BUENAPOSADA J M, BAUMELA L. Multi-task head pose estimation in-the-wild [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(8): 2874-2881. DOI:10.1109/TPAMI.2020.3046323.

[30] RUIZ N, CHONG E, REHG J M. Fine-grained head pose estimation without keypoints[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New York: IEEE Press, 2018: 2155-215509. DOI:10.1109/CVPRW.2018.00281.

[31] ZHOU Y J, GREGSON J. WHENet: Real-time fine-grained estimation for wide range head pose[C]//Proceedings of the British Machine Vision Conference 2020. Online. British Machine Vision Association, 2020: 189. DOI:10.5244/c.34.189.

[32] TAN M X, LE Q V. EfficientNet: Rethinking model scaling for convolutional neural networks[PP/OL]. *V5. arXiv* (2020-09-11). https://doi.org/10.48550/arXiv.1905.11946.

[33] YANG T Y, CHEN Y T, LIN Y, *et al*. FSA-net: Learning fine-grained structure aggregation for head pose estimation from a single image [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2019: 1087-1096. DOI:10.1109/CVPR.2019.00118.

[34] ZHU X Y, LEI Z, LIU X M, *et al*. Face alignment across large poses: A 3D solution[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2016: 146-155. DOI: 10.1109/CVPR.2016.23.

[35] KÖSTINGER M, WOHLHART P, ROTH P M, *et al*. Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization[C]//2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). New York: IEEE Press, 2011: 2144-2151. DOI: 10.1109/ICCVW. 2011.6130513.

[36] FANELLI G, DANTONE M, GALL J, *et al*. Random forests for real time 3D face analysis [J]. *International Journal of Computer Vision*, 2013, **101**(3): 437-458. DOI:10.1007/s11263-012-0549-0.

[37] ZHANG K P, ZHANG Z P, LI Z F, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks [J]. *IEEE Signal Processing Letters*, 2016, **23**(10): 1499-1503. DOI: 10.1109/LSP. 2016.2603342.

□