DOI:10.19651/j.cnki.emt.2415563

高分辨率特征保持的头部姿态软阶段回归算法*

莫建文[1] 梁豪昌[1] 袁 华[1] 姜贵昀[1] 陈明瑶[2]
(1.桂林电子科技大学信息与通信学院 桂林 541004; 2.桂林远望智能通信科技有限公司 桂林 541004)

摘 要: 针对在头部姿态估计推理过程中由于上下采样操作而导致的姿态特征损失问题,提出了一种高分辨率特征保持的头部姿态软阶段回归算法。该算法首先利用编码器 HR-Net 对原始人脸图像进行高分辨率特征保持的多尺度特征编码,并在其卷积块中加入 TA 维度交互模块以捕获更多空间与通道之间的交互信息;然后使用解码器 SSR-Net 算法对 HR-Net 输出的不同尺度特征图进行关键参数解码和头部姿态软阶段回归,并引入了高效通道注意力 ECA 以加强特征通道间的信息交互,减少冗余特征。实验结果表明,所提算法在公开数据集 AFLW2000 和 BIWI 上均有优秀表现,其 MAE 分别降低至 4.19 和 3.00。

关键词: 头部姿态估计;高分辨率特征;软阶段回归;信息交互;TA 维度交互;ECA 注意力
中图分类号: TN911.73 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Head pose estimation based high-resolution feature maintained soft-stage regression

Mo Jianwen[1] Liang Haochang[1] Yuan Hua[1] Jiang Guiyun[1] Chen Mingyao[2]
(1. School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China;
2. Guilin Yuanwang Intelligent Communication Technology Co., Guilin 541004, China)

Abstract: Aiming at the problem of pose feature loss due to up and down sampling in the inference process of head pose estimation, a high-resolution feature maintained soft-stage regression algorithm for head pose estimation is proposed. The algorithm first utilizes the encoder HR-Net to encode multiscale features for high-resolution feature maintaining in raw face images, and TA dimension interaction module joined in its convolutional block to capture more spatial-channel interaction information. The decoder SSR-Net algorithm was then applied to decode the key parameters and soft-stage regression of head pose on the different scale features output from HR-Net, and the Efficient Channel Attention ECA is employed to enhance the information interaction between feature channels and reduce redundant features. The experimental results show that the proposed algorithm has excellent performance on both the public datasets AFLW2000 and BIWI, and its MAE is reduced to 4.19 and 3.00, respectively.

Keywords: head pose estimation; high resolution feature; soft-stage regression; information interaction; TA dimension interaction module; ECA attention

0 引 言

头部姿态估计任务(head pose estimation,HPE)是通过一幅人脸数字图像来获得头部相对于相机的旋转角度。在三维空间中可以由三个欧拉角来表示物体的旋转,分别是偏转角(Yaw)、俯仰角(Pitch)和滚转角(Roll)。头部姿态估计是分析人类行为的一个重要方法,且有着非常广泛的应用,其中包括视线估计、驾驶安全与辅助、虚拟/增强现实和人机交互等[1]。

头部姿态估计任务作为计算机视觉领域的经典任务之一,在过去的数十年中受到了众多科研人员的广泛关注与深入研究。现阶段头部姿态可以划分为两大类,其中一类是传统的基于人脸关键点检测的头部姿态估计方法,另一类是基于无关键点的头部姿态估计方法。因此,本文将基于这两类方法对其相关工作进行介绍。

传统的基于人脸关键检测的头部姿态估计方法基本原理是通过检测到的人脸关键点来表示人脸,再利用关键点之间的二维空间与三维空间的转换关系来计算头部姿态欧

收稿日期:2024-03-01
*基金项目:国家自然科学基金(62177012,62001133)、广西科技重大专项(桂科 AA20302001)资助

拉角。其中 Dlib[2]使用回归树集合直接估计人脸的关键点坐标并解决了单个图像的人脸对齐问题。胡佳辉等[3]提出了利用 Dlib 进行人脸关键点定位得到眼睛图像进而进行 PnP 解算得到头部姿态信息的方法。KEPLER[4]能捕获信息丰富的结构化全局与局部特征,从而能更好地预测人脸关键点。张堃等[5]提出了一种基于关键点检测和面相人机协作系统的上肢姿态精准识别算法,能有效解决遮挡和干扰问题。EVA-GCN[6]构建了一个关键点连接图,提出利用图卷积神经网络对图类型和头部姿态欧拉角之间的复杂非线性映射进行建模。以上基于关键点的方法首先需要检测面部的关键点,随后通过建立这些关键点与 3D 头部模型之间的对应关系来恢复 3D 头部姿势,虽然这些方法可以产生较为准确的结果,但它们高度依赖关键点位置的正确预测,因此,由遮挡和极端旋转引起的较差的关键点定位会极大地降低头部姿态估计的准确率[7-8]。

基于深度学习的无关键点头部姿态估计方法为解决以上问题应运而生。其中 Ruiz 等[9]提出一种里程碑式的多重损失训练方法,通过将目标角度范围进行分组,结合交叉熵和均方误差损失函数进行欧拉角预测,给后续研究提供了一种将分类和回归相结合的思想方法。章毅等[10]使用了参数化头部姿态估计模型来避免直接回归关键点带来的歧义。FSA-Net[11]采用年龄估计领域中的软阶段回归算法,提出了一个软阶段回归结合特征聚合的网络用于欧拉角预测。TriNet[12]也采用了以上思想方法,但估计的是旋转矩阵的 3 个单位向量而非欧拉角,并加入了额外的正交损失以稳定预测结果。WHENet[13]提出了一个端到端的头部姿态估计模型,可以从单个 RGB 图像中预测整个范围头部偏转的欧拉角。LwPosr[14]首次将 Transformer 结合深度可分离卷积应用于头部姿态细粒度回归,能在较少的模型参数下有效地学习头部姿态。FDN[15]提出了一种特征解耦方法,以明确学习不同头部方向的辨别特征。TokenHPE[16]提出了以多个方向 Token 明确编码基本方向区域的方法,并构建了一种 Token 引导的损失函数,指导模型学习所需的区域相似性。MSTS-Net[17]利用不同的激活函数在 FSA-Net 的模型基础上将特征融合网络增加至三分支,但是仍然存在由多次上下采样操作导致的原始图像姿态特征损失较多的问题。为此,本文创新性地提出将用于语义分割和目标检测领域的 HR-Net[18]结合用于年龄估计领域的 SSR-Net[19]算法,巧妙地构建出一种高分辨率特征保持的头部姿态软阶段回归算法(head pose estimation based high-resolution feature maintained soft-stage regression,HSR-Net)。年龄估计和头部姿态估计本质上都是小范围区间的回归任务,所以本文将 SSR-Net 应用于头部姿态估计任务在理论上是可行的。SSR-Net 的核心思想是分 3 个阶段进行从粗到细的回归,其中每个阶段都需要 1 个特征用于预测回归参数。而 HR-Net 常用于语义分割等空间位置敏感的任务,且至多输出 4 个不同尺度的特征图,因此高度契合 SSR-Net 的核心思想。本文的主要贡献有以下几点:

1)采用 HR-Net 对原始人脸图像进行高分辨率特征保持的多尺度特征编码,结合 SSR-Net 算法对不同尺度的特征进行解码和头部姿态软阶段回归。

2)在 HR-Net 中引入 TA 维度交互模块[20],以捕获特征图中更多空间与通道之间的交互信息,聚合更多的有效特征。

3)在 SSR-Net 中引入了 ECA 高效通道注意力机制[21],减少 HR-Net 输出特征图中的冗余特征,强化通道间的交互信息。

1 HSR-Net 算法

HSR-Net 算法总体结构主要由四部分组成,其中包括主干特征提取网络 HR-Net、软阶段回归 SSR-Net 算法、轻量化 TA 维度交互模块和 ECA 高效通道注意力模块。

主干特征提取网络 HR-Net 通过并行多个分辨率分支,加上不断进行不同分支之间的信息交互,将原始图像进行高分辨特征保持的多尺度特征编码,此外还在其卷积块中加入了 TA 维度交互模块,尽可能地捕获空间与通道之间的交互信息,然后巧妙地结合软阶段回归算法 SSR-Net 对不同尺度的特征进行关键参数解码和头部姿态回归,并在 SSR-Net 中加入了 ECA 注意力机制,减少 HR-Net 输出特征中的冗余信息,强调通道间的信息交互。HSR-Net 的总体网络框架如图 1 所示。

1.1 HR-Net

由于 SSR-Net 的特殊性,分别用于从粗到细回归的阶段特征应该具备不同的尺度,即以富含头部整体信息的低分辨率特征作为第 1 阶段($k=1$)粗粒度特征,以富含局部细节信息的高分辨特征作为第 3 阶段($k=3$)细粒度特征。在 FSA-Net 和 MSTS-Net 的网络模型中分别采用双支流和三支流特征提取方式对输入的人脸图片数据进行头部姿态特征提取,虽然可以通过改变特征提取支流数量的方式来提升模型的头部姿态估计准确率,但是其输出的三个特征图分辨率相等,不具备满足 SSR-Net 特殊性的多尺度特征信息。因此多支流特征提取方式仍然存在有效特征提取能力不足的问题。

为解决以上问题和顺应 SSR-Net 的核心思想,本文采用了高分辨特征保持的 HR-Net 作为特征提取网络。HR-Net 不同于普通特征提取网络的串联卷积结构[22],而是采用了在深度和尺度两个维度上进行的并联卷积结构。如图 1 所示,在深度方向上 HR-Net 整体分为 4 个层级,每个层级又由在尺度方向上数量递增的卷积单元构成,而卷积单元主要由残差网络中不同数量的基础块(BasicBlock)或瓶颈块(Bottleneck)堆叠组成。经过前人的众多实验证明,残差网络在计算机视觉领域具备极强的特征提取能力,因此 HR-Net 中存在的大量残差结构能够对原始图像数据的

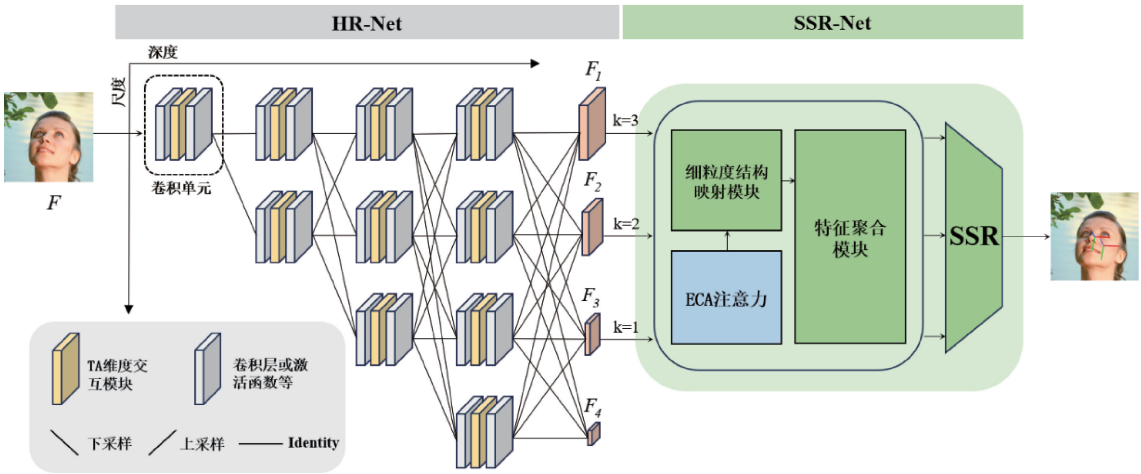

图 1　HSR-Net 的总体网络框架

姿态特征信息进行精确且鲁棒地提取。引入 TA 注意力的基础块和瓶颈块具体构成如式 (1)所示。

$$\{Conv^{3\times3}-BN-ReLU-Conv^{3\times3}-BN-TA-Residual-ReLU\}^{Basic}$$

$$\{Conv^{1\times1}-BN-Conv^{3\times3}-BN-Conv^{1\times1}-BN-TA-Residual-ReLU\}^{Bottle} \tag{1}$$

值得注意的是，HR-Net 通过上下采样和特征融合实现不同分辨率特征之间的信息交互，增强了模型对于空间和姿态信息的敏感度。其中，上采样使用最近邻采样方法(nearest neighbor interpolation，NNI)提高特征图的分辨率，并采用卷积核大小为 1×1 的卷积层将特征通道进行统一。下采样使用的并非普通的池化方式，而是采用可学习的卷积核大小为 3×3 的卷积层降低特征图的分辨率，以减少在降维过程中有效信息的损耗。上下采样的具体构成如式 (2)所示。

$$\{Conv^{1\times1}-BN-NNI\}^{Upsanple}$$
$$\{Conv^{3\times3}-BN-ReLU\}^{Downsanple} \tag{2}$$

特征融合方式采用的是按位相加，该方式能尽可能地将不同分辨率特征图中的高级语义信息和低级语义信息相结合，极大地保留多个原始特征图中的特征信息。

假定输入的原始人脸图像为 $F\in R^{3\times64\times64}$，经过 HR-Net 编码后输出至多 4 个不同分辨率的特征图 F_1、F_2、$F3_{和}$ F_4，如式 (3)所示。

$$\begin{bmatrix} F_1\in R^{32\times16\times16} & F_2\in R^{64\times8\times8} \\ F_3\in R^{128\times4\times4} & F_4\in R^{256\times2\times2} \end{bmatrix} = HR-Net(F) \tag{3}$$

1.2　TA 维度交互模块

原始人脸图像在 HR-Net 的上下采样操作中会经过大量的维度变化，这会导致姿态信息发生不可忽视的损耗。因此本文采用了一种通过三分支结构捕获维度交互信息来计算注意力权重的 TA 维度交互模块，以缓解特征信息在上下采样过程中的损失。TA 维度交互模块如图 2 所示。

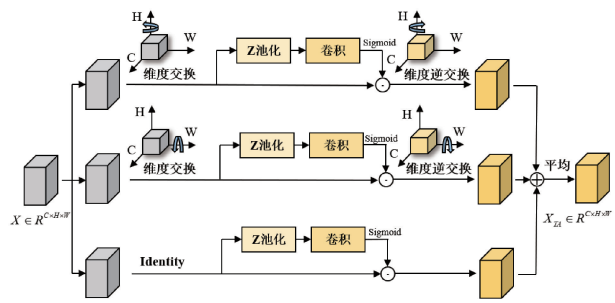

图 2　TA 维度交互模块

其中，Z 池化负责将第 0 维度方向的张量缩减至 2 维，即把第 0 维度上的平均池化特征和最大池化特征拼接起来，这既能保留原始特征的整体轮廓低频信息，也能保留局部细节的高频信息，有利于区分头部姿态数据集中细粒度的类间差距。给定一个输入张量 $X\in R^{C\times H\times W}$，则 Z 池化如式 (4)所示。

$$Z(X)=[MaxPool_{0d}(X),AvgPool_{0d}(X)] \tag{4}$$

在第一个分支中，输入张量 X 首先沿着 H 轴逆时针旋转 90°将 C 维度和 W 维度进行交换得到旋转张量 $X_1\in R^{W\times H\times C}$，再经过 Z 池化后获得张量 $X_1^Z\in R^{2\times H\times C}$，紧接着通过卷积核大小为 7×7 的卷积层 $Conv^{7\times7}$ 和 $Sigmoid$ 激活函数来生成注意力权值 $\hat{X}_1^Z\in R^{1\times H\times C}$，最后将 X_1 进行权值重分配和维度逆交换得到 $X_1^{TA}\in R^{C\times H\times W}$。以上操作能使模型在 H 维度和 C 维度之间建立信息交互，减少姿态信息的损失。第 2 个分支则是把 C 维度和 H 维度进行交换，再通过与第 1 个分支相同的操作获得 X_2^{TA}。需要注意的是，第 3 个分支不发生维度交换，而是与常见的空间注意力相同[23]，直接建立 H 维度和 W 维度之间的信息交互，然后获得 X_3^{TA}，具体流程如式 (5)所示。

$$\begin{cases} X_1^{TA}=pm^{C\leftrightarrow H}(X\odot\sigma(Conv^{7\times7}(Z(pm^{H\leftrightarrow C}(X))))) \\ X_2^{TA}=pm^{C\leftrightarrow W}(X\odot\sigma(Conv^{7\times7}(Z(pm^{W\leftrightarrow C}(X))))) \\ X_3^{TA}=X\odot\sigma(Conv^{7\times7}(Z(X))) \end{cases} \tag{5}$$

式中:pm 代表维度交换函数,σ 代表 $Sigmoid$ 激活函数。TA 注意力机制最终采用将 3 个特征按位相加取平均的方式输出最后的特征 X_{TA},如式 (6)所示。

$$X_{TA} = Avg(X_1^{TA} \oplus X_2^{TA} \oplus X_3^{TA}) \tag{6}$$

TA 维度交互模块通过跨维度交互捕获输入特征的空间维度和通道维度之间的交互信息,能够缓解姿态特征信息在上下采样过程中的损失。

1.3 SSR-Net

SSR-Net 中主要由 4 个部分组成,其中包括细粒度结构映射模块、ECA 注意力模块、特征聚合模块和 SSR 回归算法。本文改进的细粒度结构映射模块如图 3 所示。

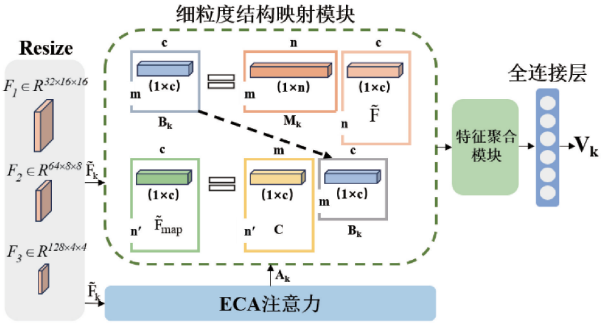

图 3 细粒度结构映射模块

为了聚合更具姿态信息的特征图和减少冗余特征,SSR-Net 采用了细粒度结构映射和特征聚合来对输入特征进行降维。输入特征图 F_k 首先被 Resize 成相同大小的特征图 $\tilde{F}_k \in R^{c\times h\times w}$,因为此时的通道维度占比大且包含大量的特征信息,所以本文引入了 ECA 高效通道注意力机制以衡量输入特征图 $\tilde{F}_k$ 中每个通道的重要性,图 3 中 $A_k = ECA(\tilde{F}_k)$,且 $A_k \in R^{c\times1\times1}$;然后将全部 $\tilde{F}_k$ 展平成二维矩阵 $\tilde{F} \in R^{c\times n}$,其中 $n = h\times w\times k$,此时 $\tilde{F}$ 中包含三个阶段全部特征。最后,特征 $\tilde{F}$ 经过单阶段映射矩阵 M_k 和跨阶段映射矩阵 C 进行结构映射后得到映射特征 $\tilde{F}_{map} \in R^{c\times n}$,具体过程如式(7)所示。

$$\begin{cases} M_k = \sigma(f_M(A_k)) \\ C = \sigma(f_C(A)) \\ \tilde{F}_{map} = M_k \otimes C \otimes \tilde{F} \end{cases} \tag{7}$$

其中,f_M 和 f_C 为全连接层,A 是由所有的 A_k 按照通道维度方向进行拼接的整体特征图,单阶段映射矩阵 M_k 计算每个 $\tilde{F}_k$ 本身像素级别的注意力权重,而跨阶段映射矩阵 C 是计算各个 $\tilde{F}_k$ 之间的注意力权重。然后将特征降维后得到的 $\tilde{F}_{map}$ 输入到特征聚合模块和全连接层输出关键参数 $V_k = \{\vec{p}^{(k)}, \vec{\eta}^{(k)}, \Delta_k\}$,再利用 SSR 算法对 V_k 进行回归,最终输出预测的头部姿态欧拉角 $\tilde{y}$。其中 SSR 算法如式 (8)所示。

$$\tilde{y} = \sum_{k=1}^{3} \vec{p}^{(k)} \cdot \vec{\mu}(k) = \sum_{k=1}^{3}\sum_{i=-1}^{1} \vec{p}(k)_i \cdot \bar{i}\left(\frac{V}{\prod_{j=1}^{k} \bar{s}_j}\right) \tag{8}$$

由于本文所使用的数据集欧拉角范围处于[-99°,99°]区间内,所以 $V=99$,$\vec{p}^{(k)}$ 代表类别概率,$\vec{\mu}^{(k)}$ 代表期望值,特别的是,期望值 $\vec{\mu}^{(k)}$ 由 $\bar{i}$ 和 $\bar{s}$ 分别采用平移因子 $\vec{\eta}^{(k)}$ 和尺度因子 Δ_k 进行动态调整,且所有 $s_k = 3$,如式 (9)所示。

$$\begin{cases} \bar{i}_k = i_k + \vec{\eta}^{(k)} \\ \bar{s}_k = s_k(1 + \Delta_k) \end{cases} \tag{9}$$

1.4 ECA 注意力

FSA-Net 中采用的是 1×1 卷积和方差函数来计算 A_k,该方法虽然能在一定程度上衡量各个通道的重要性,但是忽视了通道之间的交互信息,这在模型下游的特征降维位置中将会产生较大的姿态信息损失。为此,本文采用了 ECA 高效通道注意力机制捕获输入特征中的局部跨通道交互信息和减少冗余特征。ECA 高效通道注意力如图 4 所示。

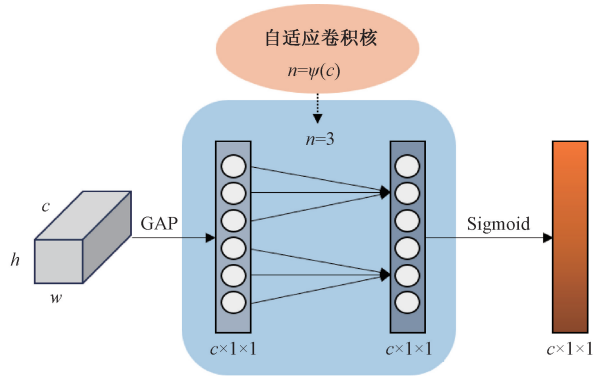

图 4 ECA 高效通道注意力

ECA 高效通道注意力使用了一个频带矩阵 W_n 去学习跨通道交互信息,如式 (10)所示。

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,n} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,n+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{c,c-n+1} & \cdots & w^{c,c} \end{bmatrix} \tag{10}$$

W_n 包含 $n\times c$ 个参数,并且避免了不同通道的完全独立。y_i 对应的权重只考虑它与它的 n 个邻居通道之间的信息交互,计算公式如式 (11)所示。

$$\omega_i = \sigma\left(\sum_{j=1}^{n} w_i^j y_i^j\right),\ y_i^j \in \Omega_i^n \tag{11}$$

$$\omega = \sigma(C1D_n(y)) \tag{12}$$

如式 (12)所示,Ω_i^n 表示 y_i 的 n 个相邻通道的集合,$C1D$ 代表一维卷积,其中的一维是指大小为 $1\times n$ 的卷积核,该卷积核代表的是跨通道信息交互的覆盖范围,且与通道数 c 成正比,其映射关系如式 (13)所示。

$$n = \psi(c) = \left| \frac{\log_2(c)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{13}$$

其中，$|t|_{odd}$ 表示取最接近 t 的奇数，在本文中 γ 和 b 分别设置为 2 和 1。通过式 (13)的映射关系，可以使高维通道具有更长的相互作用，而低维通道通过用非线性映射进行更短的相互作用，以实现卷积核 n 的自适应调整。

2　实验结果与分析

本章将详细介绍具体实验的操作细节与结果，包括实验的硬件设备条件、数据集、损失函数和与其他模型算法的对比结果等，以佐证本文提出的 HSR-Net 算法的有效性。

2.1　实验细节

本文的实验细节设置如表 1 所示。

表 1　实验细节设置

名称	设置
操作系统	Window 10
CPU	Intel i9-10900X@3.70 GHz
运行内存	64 GB
GPU	NVIDIA GeForce GTX2080Ti
深度学习框架	Pytorch 1.8.1
Python 版本	3.8
优化器	AdamW
批次大小	64
训练次数	90 epochs
初始学习率	0.001(每 30epoch 衰减 10 倍)

2.2　数据集与评价指标

1)数据集

本文使用的公开数据集包括 300W-LP、AFLW2000 和 BIWI，以下对其进行简单介绍：

(1)300W-LP 数据集

300W-LP 数据集是一个用于人脸关键点定位和姿态估计的数据集，来源于多个不同的数据集，包括 LFPW、HELEN、AFLW 和 LFW 等，涵盖了各种不同种族、年龄和性别的人脸。该数据集包含超过 12 000 张带有人脸的图像，每张图像都提供了 68 个关键点的标注和人脸姿态欧拉角标注。

(2)AFLW2000 数据集

AFLW2000 数据集由 AFLW 数据集前 2 000 张图像及其三维信息组成，三维信息由 3DMM 重建得到，因为由三维人脸模型重新标注，所以其头部姿态欧拉角有较高的精确度，常用于测试集。

(3)BIWI 数据集

BIWI 数据集由 15 000 张图像组成，邀请了 20 位志愿者参加了数据集的收集，包括 6 名女性和 14 名男性。其数据的每一帧都有一幅深度图像、RGB 图像以及头部姿态欧拉角标注。本文将 BIWI 数据集划分 70%用于训练，30%用于测试。

2)损失函数

本文采用平均绝对误差(mean absolute error，MAE)作为模型的损失函数和评价指标。损失函数 $Loss$ 如式 (14)所示。

$$Loss = L1(\tilde{y}, y) = \frac{1}{M}\sum_{m=1}^{M} \| \tilde{y}_m - y_m \| \tag{14}$$

其中，M 代表输入图像的总数量，$\tilde{y}$ 代表由模型预测的欧拉角，y 代表欧拉角的真实标签(Ground Truth)。

2.3　实验结果对比

1)消融实验

本小节将对 HR-Net、TA 维度交互模块和 ECA 注意力模块进行消融实验，该实验采用 300W-LP 数据集进行训练和 AFLW2000 数据集进行测试，具体实验设置和实验结果如表 2 所示。

表 2　消融实验

Two stream	HR-Net	TA	ECA	Yaw	Pitch	Roll	MAE
√	—	—	—	4.50	6.08	4.64	5.07
—	√	—	—	3.74	5.51	3.76	4.33
—	√	√	—	3.67	5.47	3.66	4.27
—	√	—	√	3.62	5.45	3.59	4.22
—	√	√	√	**3.61**	**5.42**	**3.54**	**4.19**

表 2 中“√”表示使用该模块，而“—”则表示不使用该模块。第一行中 Two stream 代表本文的基线模型 FSA-Net 的双流特征提取网络，在将该特征提取网络替换成 HR-Net 并引入了 TA 维度交互模块和 ECA 注意力之后 MAE 有了明显的下降，由此可知本文使用的特征提取网络结合注意力模块相比基线模型有了显著的性能提升。

为探究本文提出的 HR-Net 特征提取网络结合 TA 维度交互模块和 ECA 注意力的方法对于姿态特征提取的有效性，本文将对数据集中部分图片进行特征图展示，如图 5 所示，每个特征图是由对应特征提取网络输出且将其所有通道按位相加后的结果，其中单数行和双数行分别是 HSR-Net 和 FSA-Net 对应不同阶段输出的特征图。可以明显看出，FSA-Net 的特征图缺少人脸姿态信息，不能有效提取关键特征，而 HSR-Net 能够准确地关注到人脸的五官，例如眼睛和鼻子等，每一个阶段的特征图都蕴含着更多且精确的特征信息，这对于模型下游的头部姿态解码是非常有利的。

2)特征选取对比实验

由于 HR-Net 至多输出 4 个不同分辨率的特征图，而 SSR-Net 仅使用其中 3 个特征图就可以稳定地完成头部姿态解码，所以该小节将从 4 个特征图中选出 3 个特征图进

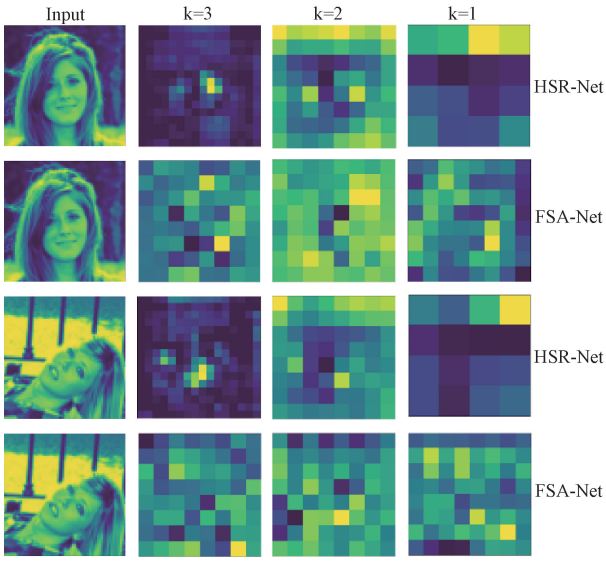

图 5 不同模型在不同阶段的特征图对比

行组合,并对其进行相应的实验,以探索特征选取的最优解。实验设置以 300W-LP 数据集进行训练,AFLW2000 数据集进行测试,实验结果如表 3 所示。

表 3 特征选取对比实验

特征选取	Yaw	Pitch	Roll	MAE
$F_1F_2F_3F_4$	**3.53**	7.58	3.92	5.01
$F_1F_2F_4$	3.66	5.52	3.56	4.25
$F_2F_3F_4$	3.67	5.47	3.66	4.27
$F_1F_3F_4$	3.70	5.55	3.61	4.29
$F_1F_2F_3$	3.61	**5.42**	**3.54**	**4.19**

表 3 中第一行代表回归的阶段数量为 4,即全部的特征都会用于 SSR-Net 头部姿态回归,其他的回归阶段数量均为 3。由表 3 可知,虽然当选取全部特征进行回归时偏转角有着不错的表现,但是俯仰角和滚转角误差较大,导致整体的 MAE 过高,而回归阶段数量取 3 时模型效果显然更佳且取前三个分辨率较高的特征时 MAE 最低,HSR-Net 此时也达到了最优性能。

3)算法对比实验

为了探究本文提出的 HSR-Net 算法的性能优劣,在本小节中将与经典头部姿态估计算法和最新的头部姿态估计算法进行对比实验。具体实验设置遵循以下协议:

(1)协议 1:在 300W-LP 数据集上进行模型训练,在 AFLW2000 数据集上进行测试,且仅考虑欧拉角范围[−99°,99°]内的数据。

(2)协议 2:在 BIWI 训练集上进行模型训练,在 BIWI 测试集上进行测试,且仅考虑欧拉角范围[−99°,99°]内的数据。

对比实验结果如表 4 所示,基于关键点检测的头部姿态估计方法在协议 1 下效果不及基于无关键点检测方法,说明本文选择的分类与回归方法这一方向的正确性。HSR-Net 在众多基于无关键点检测的头部姿态估计方法中准确性最高,且 3 个欧拉角的平均绝对误差也达到了最低,证明了本文提出的高分辨率特征保持的头部姿态回归方法拥有优秀的性能。

图 6 展示的是在协议 2 下的各个模型的欧拉角误差对比,其中蓝色虚线代表本文所提出的 HSR-Net 头部姿态估计算法,虽然 HSR-Net 的各个欧拉角 MAE 没有达到最优,但是在众多头部姿态估计方法中性能仍然属于较好水平。

表 4 算法对比实验

算法		AFLW2000				BIWI			
		Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
基于关键点	Dlib[2]	23.10	13.60	10.50	15.80	16.80	13.80	6.19	12.20
	KEPLER[4]	—	—	—	—	8.80	17.30	16.20	13.90
		协议 1				协议 2			
	EVA-GCN[6]	4.46	5.34	4.11	4.64	**2.01**	2.82	**1.89**	2.24
	Hopenet[9]	6.47	6.56	5.44	6.16	3.29	3.39	3.00	3.23
	FSA-Net[11]	4.50	6.08	4.64	5.07	2.89	4.29	3.60	3.60
	TriNet[12]	4.04	5.77	4.20	4.67	2.93	3.04	2.44	2.80
	WHENet[13]	5.11	6.24	4.92	5.42	—	—	—	—
基于无关键点	LwPosr[14]	4.80	6.38	4.88	5.35	3.62	4.65	3.78	4.01
	FDN[15]	3.78	5.61	3.88	4.42	3.00	3.98	2.88	3.29
	TokenHPE[16]	5.54	4.36	4.08	4.66	3.01	**2.28**	2.01	2.49
	MSTS-Net[17]	4.22	5.62	4.20	4.68	—	—	—	3.59
	HSR-Net (ours)	**3.61**	**5.42**	**3.54**	**4.19**	2.91	3.70	2.39	3.00

协议 1 和协议 2 模型训练过程曲线分别如图 7(a)和(b)所示。可知在协议 1 和协议 2 下训练的模型其训练

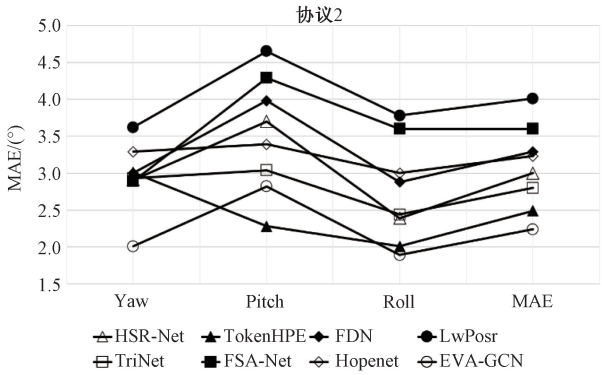

图 6　协议 2 下的欧拉角误差对比

损失整体从高到低逐渐收敛，且随着学习率衰减在每 30 epochs 的节点有明显的下降，测试损失也随着训练损失的下降而整体稳步下降，在最后的 30 epochs 阶段基本达到收敛状态。

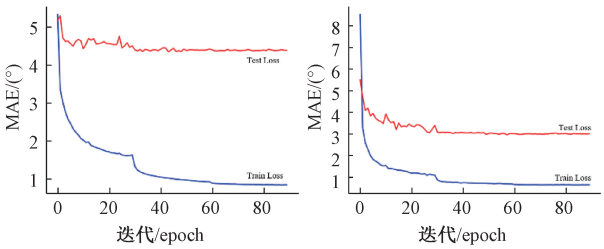

图 7　训练过程曲线

4)实验结果对比

为了能更加直观地观察 HSR-Net 算法的有效性和探索其实际应用的性能，本小节将在 AFLW2000 测试集中选取部分不同姿态的数据进行结果展示，如图 8 所示。其中单数行和双数行分别代表 HSR-Net 和 FSA-Net 的预测结果，图中蓝色线表示人脸的俯仰角，红色线表示人脸的偏转角，绿色线表示人脸的滚转角，此外，图片中心的三根线表示由模型预测的结果，而偏右下方的三根线表示姿态欧拉角真实标签。

图 8　模型预测结果对比

由图 8 可知，虽然两个模型单纯从肉眼看基本都能预测出一个较为准确的结果，但是通过仔细观察可以发现其中的细节区别，由本文提出的 HSR-Net 算法在各个欧拉角中都有更加接近真实标签的预测结果，例如在图 8 中的最后一张人脸上 HSR-Net 预测的俯仰角显然比 FSA-Net 预测的更加精准，这说明了本文提出的 HSR-Net 在实际应用中有着优秀的性能。

3　结　　论

本文采用的是基于无关键点头部姿态估计方法中的基于分类与回归的方法，该方法需要采用深度卷积神经网络进行原始人脸图像的特征提取，其中涉及的大量上下采样操作会导致图像中的姿态信息严重损失，因此，本文提出了高分辨率特征保持的 HR-Net 结合头部姿态软阶段回归算法 SSR-Net 的新方法，并引入了加强空间和通道信息交互的 TA 维度交互模块和 ECA 高效通道注意力机制，强调不同尺度、空间和通道之间的特征交互，极大地保留原始图像中的姿态信息，使模型对姿态和空间更加敏感。实验表明，本文提出的 HSR-Net 头部姿态估计算法在不同的公开测试数据集中均有优秀的表现，尤其在难度较高的 AFLW2000 数据集中 MAE 达到了 4.19，此外，该算法在实际应用中的准确性高，有着较强的实际应用价值。

参考文献

[1] ASPERTI A, FILIPPINI D. Deep learning for head pose estimation: A survey[J]. SN Computer Science, 2023, 4(4): 349.

[2] KAZEMI V, SULLIVAN J. One millisecond face alignment with an ensemble of regression trees[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1867-1874.

[3] 胡佳辉，陆永华，张进海，等. 基于机器学习的头部自由视线追踪方法及其在电动病床端的应用[J]. 仪器仪表学报，2021，42(12)：101-109.

[4] KUMAR A, ALAVI A, CHELLAPPA R. KEPLER: Simultaneous estimation of keypoints and 3D pose of unconstrained faces in a unified framework by learning efficient H-CNN regressors[J]. Image and Vision Computing, 2018, 79: 49-62.

[5] 张堃，刘志诚，刘纪元，等. 面向人机协作系统的上肢姿态精准识别算法研究[J]. 仪器仪表学报，2023，44(1)：275-282.

[6] XIN M, MO S, LIN Y. Eva-gcn: Head pose estimation based on graph convolutional networks[C]. Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition,2021: 1462-1471.

[7] STOREY G, BOURIDANE A, JIANG R. Integrated deep model for face detection and landmark localization from "in the wild" images[J]. IEEE Access, 2018, 6: 74442-74452.

[8] YUEN K, TRIVEDI M M. An occluded stacked hourglass approach to facial landmark localization and occlusion estimation [J]. IEEE Transactions on Intelligent Vehicles, 2017, 2(4): 321-331.

[9] RUIZ N, CHONG E, REHG J M. Fine-grained head pose estimation without keypoints[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,2018: 2074-2083.

[10] 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/OL].软件学报,2023,1-16[2024-01-23]. https://doi.org/10.13328/j.cnki.jos.007029. DOI: 10.13328/j.cnki.jos.007029.

[11] YANG T Y, CHEN Y T, LIN Y Y, et al. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition.,2019: 1087-1096.

[12] CAO Z, CHU Z, LIU D, et al. A vector-based representation to enhance head pose estimation[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2021: 1188-1197.

[13] ZHOU Y, GREGSON J. Whenet: Real-time fine-grained estimation for wide range head pose[J]. Arxiv Preprint, 2020,Arxiv:2005.10353.

[14] DHINGRA N. Lwposr: Lightweight efficient fine grained head pose estimation[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 1495-1505.

[15] ZHANG H, WANG M, LIU Y, et al. FDN: Feature decoupling network for head pose estimation[C]. Proceedings of the AAAI Conference on Artificial Intelligence,2020, 34(7): 12789-12796.

[16] ZHANG C, LIU H, DENG Y, et al. TokenHPE: Learning orientation tokens for efficient head pose estimation via transformers[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2023: 8897-8906.

[17] 韩雪,张红英,卢琇雯,等.多阶段特征融合的三支流头部姿态估计算法[J].计算机工程与应用,2023,59(17):212-222.

[18] WANG J, SUN K, CHENG T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3349-3364.

[19] YANG T Y, HUANG Y H, LIN Y Y, et al. Ssr-net: A compact soft stagewise regression network for age estimation[C]. IJCAI,2018, 5(6): 7.

[20] MISRA D, NALAMADA T, ARASANIPALAI A U, et al. Rotate to attend: Convolutional triplet attention module[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2021: 3139-3148.

[21] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020: 11534-11542.

[22] FAN Z, ZHU Y, HE Y, et al. Deep learning on monocular object pose detection and tracking: A comprehensive overview [J]. ACM Computing Surveys, 2022, 55(4): 1-40.

[23] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

作者简介

莫建文(通信作者),副教授,硕士生导师,主要研究方向为机器视觉,智能信号处理。

E-mail:Mo_jianwen@126.com

梁豪昌,硕士研究生,主要研究方向为图像识别。

E-mail:1286025669@qq.com

袁华,讲师,硕士,主要研究方向为图像信号处理。

E-mail:yuanhua@guet.edu.cn

姜贵昀,硕士研究生,主要研究方向为图像识别。

E-mail:1060526807@qq.com

陈明瑶,学士,主要研究方向为人工智能。

E-mail:463488607@qq.com