基于多层次自注意力网络的人脸特征点检测

徐浩宸，刘满华

（上海交通大学电子信息与电气工程学院，上海 200240）

摘　要：人脸特征点检测是人脸图像处理的关键步骤之一，常用检测方法是基于深度神经网络的坐标回归方法，具有处理速度快的优点，但是用于回归的高层次网络特征丢失空间结构信息，且缺乏细粒度表征能力，导致检测精度降低。针对该问题，提出一种基于多层次自注意力网络的人脸关键点检测算法。为提取更具有细粒度表征能力的图像语义特征，构建基于自注意力机制的多层次特征融合模块，实现高层次高语义信息特征和低层次高空间信息特征的跨层次特征融合。在此基础上，设计一种多任务学习人脸特征点检测定位与人脸姿态角估计的训练方式，优化网络对人脸整体朝向姿态的估计，以提升特征点检测的准确性。在人脸特征点主流数据集 300W 和 WFLW 上的实验结果表明，与 SAAT、AnchorFace 等方法相比，该方法有效提升网络的检测精度，标准平均误差指标分别为 3.23% 和 4.55%，相较于基线模型降低 0.37 和 0.59 个百分点，在 WFLW 数据集上错误率指标为 3.56%，相较于基线模型降低了 2.86 个百分点，能够提取更具鲁棒性和细粒度的表达特征。

关键词：人脸特征点检测；卷积神经网络；自注意力机制；特征融合；多任务学习；深度学习

源代码链接：https://github.com/Cosmo1210/Hierarchical-Self-attention-Network

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0066927

Facial Landmark Detection Based on Hierarchical Self-Attention Network

XU Haochen，LIU Manhua

（School of Electronic Information and Electrical Engineering，Shanghai Jiao Tong University，Shanghai 200240，China）

【Abstract】 Facial landmark detection，a key step in facial image processing，is commonly performed using the coordinate regression method based on deep neural networks，which has the advantage of fast processing speed. However，the high-level network features used for regression lose spatial structural information and lack fine-grained representation ability，leading to a decrease in detection accuracy. Therefore，a facial landmark detection algorithm based on multi-level self-attention network is proposed to address this issue. To extract image semantic features with finer granularity representation ability，a multi-level feature fusion module based on self-attention mechanism is constructed to achieve cross-level fusion of high-level semantic and low-level spatial information features. On this basis，a training method for multi-task learning of facial landmark detection and localization，as well as facial pose angle estimation，is designed to optimize the network estimation of the overall orientation and pose of the face，thereby improving the accuracy of landmark detection. The experimental results on mainstream facial landmark datasets 300W and WFLW show that compared with methods such as SAAT and AnchorFace，the proposed method effectively improves network detection accuracy，achieving a standard average error of 3.23% and 4.55%，respectively，which are 0.37 and 0.59 percentage points lower than the baseline model. The error rate indicator on the WFLW dataset is 3.56%，which is 2.86 percentage points lower than the baseline model，demonstrating that the proposed method can extract more robust and fine-grained expression features.

【Key words】 facial landmark detection；Convolutional Neural Network（CNN）；self-attention mechanism；feature fusion；multi-task learning；deep learning

0　引言

人脸特征点检测又称人脸对齐，指自动定位人脸上的一系列预设基准点（如眼角、嘴角等），是人脸处理的关键步骤。人脸特征点检测是许多人脸相关视觉任务的基本组成部分，被广泛应用于如人脸识别、表情分析、虚拟人脸重建等领域[1]。

传统算法在环境受限条件下的人脸特征点检测可以得到较准确的结果，如主动外观模型（AAM）[2]、约束局部模型（CLM）[3]等。该领域的挑战是在非受限环境下的人脸特征点检测，在非受限环境下，人脸会具有受限环境下所没有的局部变化及全局变化。局部变化包括表情、遮挡、局部的高光或阴影等，这些局部变化使得一部分人脸特征点偏离正常位置，

收稿日期：2023-02-13　　**修回日期：**2023-03-30

基金项目：国家自然科学基金面上项目（62171283）；上海市自然科学基金（20ZR1426300）；上海市市级科技重大专项（2021SHZDZX0102）。

通信作者 E-mail：xuhaochen@sjtu.edu.cn

乃至于消失不见。全局变化包括面部的大姿态旋转、图片模糊失焦等,这些全局变化使得大部分人脸姿态点偏离正常位置。这2个挑战需要算法模型对人脸特征点的全局和局部分布有良好的表征,对形状分布有足够的鲁棒性,对人脸的姿态朝向有所估计。

研究人员尝试使用一些传统的级联回归方法解决在非受限环境下人脸特征点检测回归任务,这些方法可以归纳为级联一系列弱回归器以训练组合成1个强回归器。然而这些方法在较浅的级联深度后其性能会达到饱和,精度难以再次提高。

随着基于卷积神经网络的发展,在非受限环境下的人脸特征点检测得到极大改善。现有基于卷积神经网络的方法主要分为基于坐标回归的方法和基于热图回归的方法。基于热图回归的方法通过级联Hourglass网络[4]得到像素级的热图估计值,准确率较高,但是由于级联网络的结构和预测整个热图值,因此参数量较大且推理时间长。基于坐标回归的方法利用卷积神经网络直接回归出特征点的坐标,参数量较少且推理时间短,实时性较好。特征点坐标精确到像素,需要足够的空间信息才能保证精度。而基于坐标回归的方法模型随着网络的加深和降采样,在特征语义信息加深的同时也会丢失空间结构信息,缺乏细粒度表征能力,精度会有所降低。

本文针对人脸特征点检测坐标回归方法提出一种多层次自注意力网络(HSN)模型,构建一种基于自注意力机制的多层次特征融合模块,实现网络的跨层次特征融合,提升用于回归特征的空间结构信息,弥补细粒度表征能力不足。此外,设计一种多任务同时学习特征点检测定位及人脸姿态角估计的训练方式,提升模型对人脸姿态朝向的表征,从而提升模型的准确性。

1 相关研究

1.1 传统的人脸特征点检测方法

传统人脸特征点检测以传统算法为主,通过多训练级联回归器来构建算法。COOTES等[2]提出AAM算法,该算法根据人脸的整体外观、形状、纹理参数化建立模型,通过迭代搜索特征点位置,并应用平均人脸修正结果,最大化图像中局部区域的置信度以完成置信度检测。CRISTINACCE等[3]在AAM算法的基础上,弃用全局纹理建模方法,利用一系列特征点周围局部纹理约束模型,构建CLM算法[3]。DOLLAR等[5]提出级联姿态回归器(CPR),通过级联一系列回归器实现预测值的不断修正细化,以得到最终预测值。

1.2 基于深度学习的人脸特征点检测方法

近年来,基于深度学习的人脸特征点检测方法表现出了远优于传统算法的性能,大致可以分为基于坐标回归的方法与基于热图回归的方法。基于坐标回归的方法使用卷积神经网络直接从图片输入中回归出特征点坐标值。SUN等[6]提出一种三级级联的卷积神经网络,从粗到细地定位人脸特征点。GUO等[7]基于MobileNet结构做进一步的轻量缩减,提出一种可以在移动设备上也能实时运行的网络结构,且针对不同的特殊数据类别自适应地加权训练。FENG等[8]针对面部特征点检测这一特定任务,设计一种WingLoss损失函数,使得损失函数在大误差时的梯度为常数,在小误差时得到比L1与L2更大的误差值。ZHANG等[9]提出一种多任务学习框架,同时优化人脸特征点定位及姿态、表情、性别等面部属性分类。基于热图回归的方法为每个特征点预测1张热图,得到每个像素位置的概率置信度值,然后从热图中估计得到坐标值。受Hourglass在人体姿态估计方面的启发,Hourglass网络成为很多业界使用热图回归法的主干网络。YANG等[10]使用有监督的面部变换归一化人脸,然后使用Hourglass回归热图。DENG等[11]提出JMFA网络,通过堆叠Hourglass网络,在多视角人脸特征点检测上达到较高的精度。文献[12]提出LAB网络,利用额外的边界线描述人脸图像的几何结构,从而提升特征点检测的准确性。热图中背景像素会逐渐收敛于零值,而WingLoss函数在零点不连续,导致无法收敛。针对该问题,WANG等[13]提出一种改进的自适应WingLoss函数,使其能适应真实值热图不同的像素强度,当损失函数在零值附近时接近于L2 Loss,因此可支持热图回归训练。

1.3 卷积神经网络中的特征融合方法

卷积神经网络的底层特征包含大量的空间结构信息而缺乏语义信息,随着网络的加深和降采样,高层特征具备丰富的语义信息而丢失空间结构信息。运用单独的高层卷积网络特征对于精细任务来说是不足的。学界也有一些工作探索在1个卷积神经网络中运用不同卷积层的有效性,如特征金字塔网络(FPN)通过融合低层次的高分辨率特征与上采样后的高层次高语义特征得到不同分辨率的特征,以支持不同尺度的目标检测任务[14]。HARIHARAN等[15]尝试使用卷积神经网络中的所有特征,以提升网络在定位任务中的精度。LONG等[16]在分割任务中结合不同深度间更高层及更精细的特征。XIE等[17]在边缘检测任务中设计1个整体嵌套的网络框架,网络的旁路输出被加到较底层的卷积层后,以提供深层监督训练。

简单级联不同卷积层特征会增加大量的参数量,同时将较多中间卷积层相结合时不能捕捉层间的交互关系。受Transformer网络[18]中自注意力机制的启发,本文将每个网络块的特征视为不同的网络特征提取器,并运用自注意力机制建模层与层之间的交互融合。

2 本文方法

本文所提基于多层次自注意力网络 HSN 的人脸特征点检测算法的总体流程如图 1 所示。其中实线框为训练与测试时的通用流程与数据，虚线框为仅在训练过程中所使用的流程及数据。在算法测试时流程分为数据预处理及 HSN 模型计算 2 个阶段。对于输入图像，首先人脸识别后进行归一化处理。在模型训练阶段，对于没有人脸姿态角真值的数据集须额外计算 1 个拟真值，并在数据增强后用于网络训练。对于有遮挡、表情、局部高光或阴影等细分类的数据集，其分类真值将用于损失函数的计算。

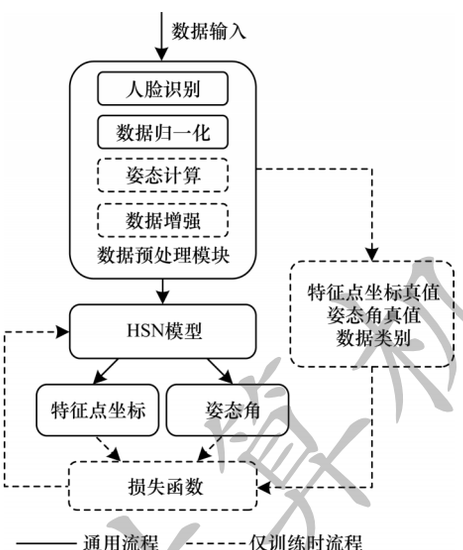

图 1 基于多层次自注意力网络的人脸特征点检测算法流程

Fig.1 Procedure of facial landmark detection algorithm based on hierarchical self-attention network

2.1 数据预处理

大多数现有人脸特征点检测方法建立在人脸识别技术的基础上，算法在经人脸识别后的输入图像局部区域上检测人脸特征点。部分数据集提供人脸区域坐标框或直接提供人脸区域局部图像。针对未提供该部分数据的数据集，本文直接采用 MTCNN 算法[19]进行人脸识别，在检测人脸区域上继续后处理。

人脸特征点示意图如图 2 所示。针对未提供人脸姿态角真值的数据集，本文提出计算 1 个拟真值用于网络训练，方法如下：将该数据集和训练集的所有正脸图像经由双眼外侧 2 点旋转至水平矫正后，选取如图 2 圆点所示的 14 个点作为基准点，统计训练集全部平均点作为该数据集的标准人脸。将人脸大致视为刚体，将每张图像的 14 个基准点相对标准人脸计算其旋转矩阵，然后计算 3 个欧拉角作为该数据集的人脸姿态角拟真值。由于训练数据涉及坐标及姿态角，因此本文数据增强仅使用了图像翻转方法。

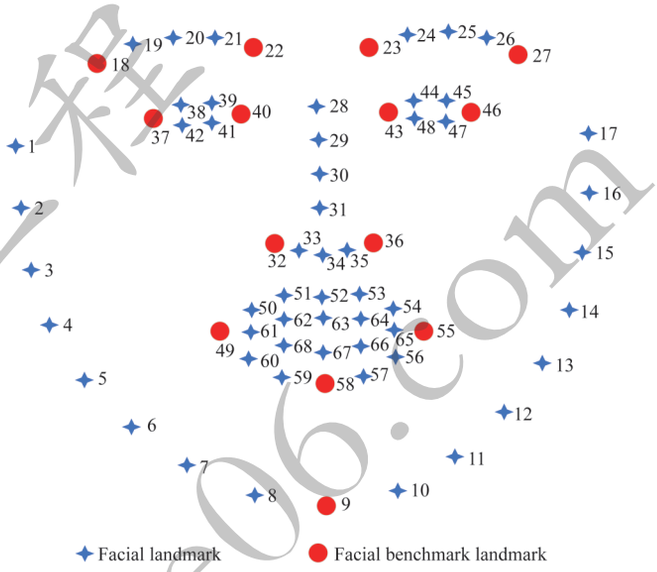

图 2 人脸特征点示意图

Fig.2 Schematic diagram of facial landmark

2.2 多层次自注意力网络

本文提到针对非受限环境下的人脸特征点检测算法模型对全局形状应具备足够的鲁棒性，在局部分布上应具备细粒度的表征能力，对人脸位姿朝向应有所估计。针对以上问题，本文提出一种基于多层次自注意力的特征融合网络，网络结构如图 3 所示。输入图像经由主干网络、多层次自注意力特征融合模块后输出预测特征点坐标及预测人脸位姿角。

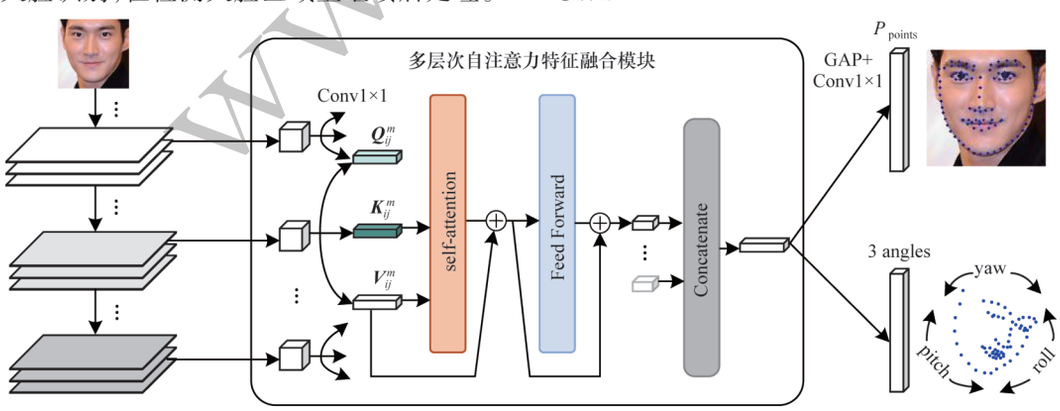

图 3 多层次自注意力网络结构

Fig.3 Structure of hierarchical self-attention network

2.2.1 主干网络

HSN 模型采用 ResNet 50[20-21]作为主干网络，其结构如图 4 所示，由卷积层、池化层及瓶颈卷积块组成。主干网络分为 5 个阶段来分层学习特征，每个阶段之间采用步长为 2×2 的池化层或卷积层进行降采样。第 1 个阶段由 7×7 的卷积层、BN 层及 ReLU 激活函数组成。其他 4 个阶段均由连续的瓶颈卷积块组成，每个瓶颈卷积块由 2 个用于升降维的 1×1 卷积核，1 个 3×3 的卷积核、BN 层和 ReLU 激活函数构成。5 个阶段的通道数设置为 64、256、512、1 024 和 2 048。

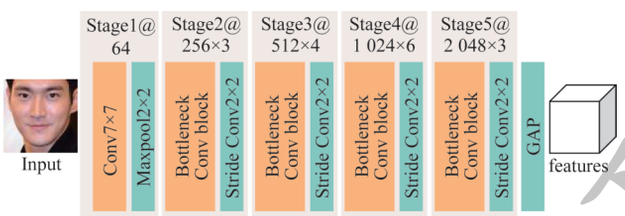

图 4 主干网络结构

Fig.4 Structure of backbone network

2.2.2 多层次自注意力特征融合模块

现有基于坐标回归的方法多直接基于主干网络顶层特征全局平均池化后回归，虽然该特征因其共享特性具有全局形状的鲁棒性，但是丢失空间结构信息，在局部细粒度方面表征不足，不足以描述局部细节的语义信息。这些方法忽略了中间层的激活，导致细粒度判别信息损失。本文将主干网络层中每个阶段的网络块都视为不同的特征提取器，将各个特征块的激活视为不同特性的响应，并通过自注意力机制捕捉层间关系，以提升模型细粒度表征能力。

假定输入图像 $\boldsymbol{I}$，经由主干网络后第 s 阶段输出高度为 h，宽度为 w，通道数为 c 的特征 $\boldsymbol{X}_s \in \mathbb{R}^{h\times w\times c}$。定义特征 $\boldsymbol{X}_s$ 中高度为 i，宽度为 j 位置处的特征向量为 $\boldsymbol{x}_{sij}=\left[x_{sij}^1, x_{sij}^2, \cdots, x_{sij}^c\right]^{\mathrm{T}}$。网络阶段数为 L，网络在不同阶段间经过步长为 2 的下采样，故 $\boldsymbol{X}_s$ 前 t 阶段对应位置的特征向量 $\boldsymbol{y}\in \boldsymbol{Y}_{S-t,ij}=\left\{\boldsymbol{x}_{(s-t)ab} \middle| a\in\left[2ti, 2t(i+1)-1\right], b\in\left[2tj, 2t(j+1)-1\right]\right\}$。以最后阶段的输出特征图相对位置为准，组合 m 个阶段对应位置的对应特征向量作为该位置的多层次特征向量 $\boldsymbol{S}_{ij}^m=\left[\boldsymbol{f}_0^{1\times1}(\boldsymbol{x}_{Lij}), \boldsymbol{f}_1^{1\times1}(\boldsymbol{y}) \middle| \boldsymbol{y}\in\boldsymbol{Y}_{L-1,ij}, \cdots, \boldsymbol{f}_{m-1}^{1\times1}(\boldsymbol{y}) \middle| \boldsymbol{y}\in\boldsymbol{Y}_{L-m+1,ij}\right]$，维数为 $v=\sum_{k=1}^{m}2^{k-1}$，$f^{1\times1}$ 为 1×1 的卷积核及相应的 BN 层，用于统一不同特征块的通道数至 $c'=\frac{c}{2^m}$，则 $\boldsymbol{S}_{ij}^m\in\mathbb{R}^{v\times c'}$。

本文所提的多层次自注意力特征融合模块如图 3 所示。3 个 1×1 的卷积核分别用于将输入特征 $\boldsymbol{S}_{ij}^m$ 转换成查询向量、索引向量、内容向量并组合为矩阵形式 $\boldsymbol{Q}_{ij}^m, \boldsymbol{K}_{ij}^m, \boldsymbol{V}_{ij}^m\in\mathbb{R}^{v\times c'}$，然后利用自注意力层对多层次特征向量建模，计算查询向量和索引向量间的内积并归一化作为相关系数，经 Softmax 激活后加权内容向量，获得表征 $\boldsymbol{R}_{ij}^m$：

$$\boldsymbol{R}_{ij}^m=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{ij}^m\left(\boldsymbol{K}_{ij}^m\right)^{\mathrm{T}}}{\sqrt{c'}}\right)\boldsymbol{V}_{ij}^m \tag{1}$$

输出表征 $\boldsymbol{R}_{ij}^m$ 展平并经过 2 层前馈神经网络后连接得到最终输出特征。自注意力层及前馈神经网络均有跳跃连接结构。上述结构在不同位置处共享权重，输出特征经由全局平均池化后连接输出层。

2.2.3 损失函数

为优化网络对人脸整体朝向姿态的估计，以提升特征点定位的准确性，本文网络在输出预测特征点坐标的同时输出预测人脸姿态角，并将其加入到损失函数中以优化训练。损失函数由特征点坐标的损失函数与位姿角的损失函数加权得到：

$$L=\frac{1}{N}\sum_{n=1}^{N}\sum_{u=1}^{U}\omega_u\left(\mathcal{L}_\theta+\alpha\mathcal{L}_d\right) \tag{2}$$

其中：N 为每次迭代训练图像数；U 为图像细分类类别数量；ω_u 为各细分类权重，本文采用各类别图像占比的倒数加权，以增加模型对稀少数据的敏感性，即 $\omega_u=N_t/N_u$，N_t 为训练集图像总数，N_u 为训练集第 u 类图像总数；$\mathcal{L}_\theta$ 和 $\mathcal{L}_d$ 分别为位姿角和特征点坐标的损失函数；α 为平衡 2 项损失函数的超参数。本文中位姿角损失函数 $\mathcal{L}_\theta$ 定义如下：

$$\mathcal{L}_\theta=\sum_{z=1}^{3}\Delta\theta_z^{\,2} \tag{3}$$

其中：$\Delta\theta_z$ 为人脸的 3 个姿态角预测值与真值的差值，$z\in\{1,2,3\}$。

本文特征点损失函数 $\mathcal{L}_d$ 选用 WingLoss[8]：

$$\mathcal{L}_d=\sum_{p=1}^{P}\begin{cases}\sigma\ln\left(1+\dfrac{\left\|\Delta d_p\right\|_2}{\epsilon}\right), & \left\|\Delta d_p\right\|_2<\sigma\\ \left\|\Delta d_p\right\|_2-\phi, & \text{其他}\end{cases} \tag{4}$$

其中：P 为预测特征点数；$\left\|\Delta d_p\right\|_2$ 为特征点 p 坐标预测值与真值的 L2 距离；σ,ϵ 为超参数；ϕ 为保证曲线连续的常数，$\phi=\sigma-\sigma\ln\left(1+\dfrac{|\sigma|}{\epsilon}\right)$。

3 实验结果与分析

为测试本文所提出方法的性能，采用人脸特征点检测领域最常用的 2 个数据集 300W 数据集[22]以及 WFLW 数据集[12]上进行包括消融实验在内的一系列实验。

3.1 数据集

300W 数据集重新标定了包括 XM2VTS、FRGC Ver.2、

LFPW、HELEN、AFW、iBUG 6个人脸数据集的人脸特征点数据，提供68个人脸特征点坐标，在其分支数据集300W-LP中提供人脸位姿角数据。本文遵照前人方法的实验设置[12]，取用来自LFPW、HELEN、AFW数据集总共3 148张图像作为训练集，来自LFPW和HELEN的测试集以及iBUG总共689张图像作为测试集图像。测试图像包含普通图像及有挑战图像2大类。由于300W数据集不包含细分类数据，因此损失函数中细分类权重统一为1。

WFLW数据集包括7 500张图像的训练集及2 500张图像的测试集，提供98个人脸特征点坐标以及包括大姿态角、夸张表情、极端光照、化妆、遮挡、模糊6项细分类数据。在本文方法中增加第7类普通类，包含所有不在其他所有类中的数据，以便细分类权重的计算。

3.2 实验设置与测试指标

本文模型的搭建与训练，数据集的测试均在PyTorch框架下进行。本文中特征融合模块选择融合最后3个阶段的特征，损失函数平衡项α取值为1，特征点损失函数WingLoss中σ取值为10，ϵ取值为2。本文训练与测试时Batch_size设置为96，使用AdamW优化器优化训练，学习率设置遵照5×10^{-7}~1×10^{-5}的余弦退火策略，模型训练迭代次数Epoch设为300。

为方便与现有方法进行对比，本文使用标准化平均误差（NME，计算中用N_{NME}）以及错误率（FR，计算中用F_{FR}）衡量性能。本文的NME指标按照双眼外眼角间距离进行归一化。NME和FR如式（5）和式（6）所示：

$$N_{\mathrm{NME}}=\frac{1}{N}\frac{1}{P}\sum_{n=1}^{N}\sum_{p=1}^{P}\frac{\left\|Y_p-\widehat{Y}_p\right\|}{d}\times100\% \tag{5}$$

$$F_{\mathrm{FR}}=\frac{1}{N}\sum_{n=1}^{N}\left[N_{\mathrm{NME}_n}\geqslant10\%\right]\times100\% \tag{6}$$

其中：N为测试图像数；P为预测特征点数；Y_p和$\widehat{Y}_p$分别为人脸特征点坐标的真值以及预测值；d为归一化因子；N_{NME_n}为具体每张测试图像的NME值。

3.3 与现有方法的比较

在300W数据集上测试本文所提方法HSN的性能，并与现有方法进行对比，实验结果如表1所示，加粗表示最优数据。表1中AnchorFace[23]、Wing、PFLD为基于坐标回归的方法，LAB[12]、MobileFAN[24]、HRNetV2[25]为基于热图回归的方法，LDDMM-Face[26]是基于形状模型的其他方法。从表1可以看出，本文所提的HSN模型在基于坐标回归方法中表现出优异的性能，NME降至3.23%。在有挑战图像大类中，本文模型的NME降至5.12%，验证HSN模型补充坐标回归方法特征细粒度表征不足的假设。在与HRNetV2等热图回归方法、LDDMM-Face等形状模型对比时，本文模型在各指标对比上也具有一定的优越性。

表1 在300W数据集上不同模型的标准化平均误差对比

Table 1 Normalized mean error comparison among different models on the 300W dataset %

模型	标准化平均误差		
	普通图像	有挑战图像	全部图像
AnchorFace	3.12	6.19	3.72
Wing	3.01	6.01	3.60
PFLD	3.32	6.56	3.92
LAB	3.42	6.98	4.12
MobileFAN	2.98	5.34	3.45
HRNetV2	2.87	5.15	3.32
LDDMM-Face	3.07	5.40	3.53
HSN	**2.77**	**5.12**	**3.23**

同样地，在WFLW数据集上本文模型与现有模型进行一系列对比，结果如表2所示。从表2可以看出，本文模型HSN在NME及FR 2个指标上均取得更高精度，分别为4.55%和3.56%。在各个困难细分类中，本文模型在大姿态角及模糊2类上的NME分别降到了7.76%和5.17%，验证本文所提的加入姿态角预测辅助训练以及跨层融合方法的有效性和先进性。

表2 在WFLW数据集上不同模型的评价指标

Table 2 Evaluation indicators among different models on the WFLW dataset %

模型	测试集		细分类测试集NME					
	NME	FR	大姿态角	夸张表情	极端光照	化妆	遮挡	模糊
LAB	5.27	7.56	10.24	5.51	5.23	5.15	6.79	6.32
Wing	5.11	6.00	8.75	5.36	4.93	5.41	6.37	5.81
SAAT	5.11	5.63	—	—	—	—	—	—
MobileFAN	4.93	5.32	8.72	5.27	4.93	4.70	5.94	5.73
HRNetV2	4.60	—	7.94	4.85	4.55	**4.29**	**5.44**	5.42
LDDMM-Face	4.63	3.68	—	—	—	—	—	—
AnchorFace	4.62	4.20	—	—	—	—	—	—
HSN	**4.55**	**3.56**	**7.76**	**4.85**	**4.45**	4.49	5.47	**5.17**

3.4 消融实验

为进一步验证HSN的有效性，本文在WFLW数据集上进行一系列消融实验。

本文基于提升模型对人脸整体姿态朝向的表征能力来提高特征点定位精度的假设，设计多任务学习特征点定位及人脸姿态角的训练方式，并对细分类的困难项增加了相应的权重。损失函数消融实验结果如表3所示。该部分消融实验所用的网络均为仅使用主干网络的基准网络，不加权且仅预测特征点坐标作为基线对比。从表3可以看出，分别添加位姿角预测损失项（$\mathcal{L}_\theta$）和细分类权重（ω_u）后，HSN在特征点定位的NME及FR指标均得到有效优化，NME分别下降0.21和0.09个百分点，FR分别下降0.87及0.44个百分点。在两者综合作用下，NME和FR分别下降0.26和1.29个百分点。

表3 损失函数消融实验结果

Table 3 Results of loss function ablation experiment %

$\mathcal{L}_d$	$\mathcal{L}_\theta$	ω_u	NME	FR
√			5.14	6.42
√	√		4.93	5.55
√		√	5.05	5.98
√	√	√	**4.88**	**5.13**

为验证该特征融合模块的有效性，本文选取主干网络作为基准，融合实验结果如表4所示。首先与最常用的特征融合方法拼接法（Concatenate）与加和法（Addition）进行对比，同样是在融合第4个和第5个阶段特征的情况下，本文方法能有效改善NME值，加和法的性能不增反降，而相较于拼接法，本文方法在NME和FR 2个指标上分别下降0.11和0.55个百分点。而后对特征融合的阶段数进行实验，其中融合第3~5个阶段的模型表现出最优性能。融合第2~5个阶段的实验可能是由于第2个阶段的特征缺乏足够的语义信息而给融合后的特征带来无效噪声，导致性能略逊于第3~5个阶段的特征融合网络。

表4 特征融合消融实验结果

Table 4 Results of feature fusion ablation experiment %

方法	NME	FR
Baseline	4.88	5.13
+ $\text{Concatenate}_{4,5}$	4.79	4.91
+ $\text{Addition}_{4,5}$	5.60	8.40
+$\text{HSN}_{4,5}$	4.68	4.36
+$\text{HSN}_{3,4,5}$	**4.55**	**3.56**
+$\text{HSN}_{2,3,4,5}$	4.56	3.72

3.5 可视化结果与讨论

为进一步直观展示HSN的有效性，本文将HSN和其他方法在WFLW数据集中的不同困难细分类测试集上的测试结果可视化，结果如图5所示。

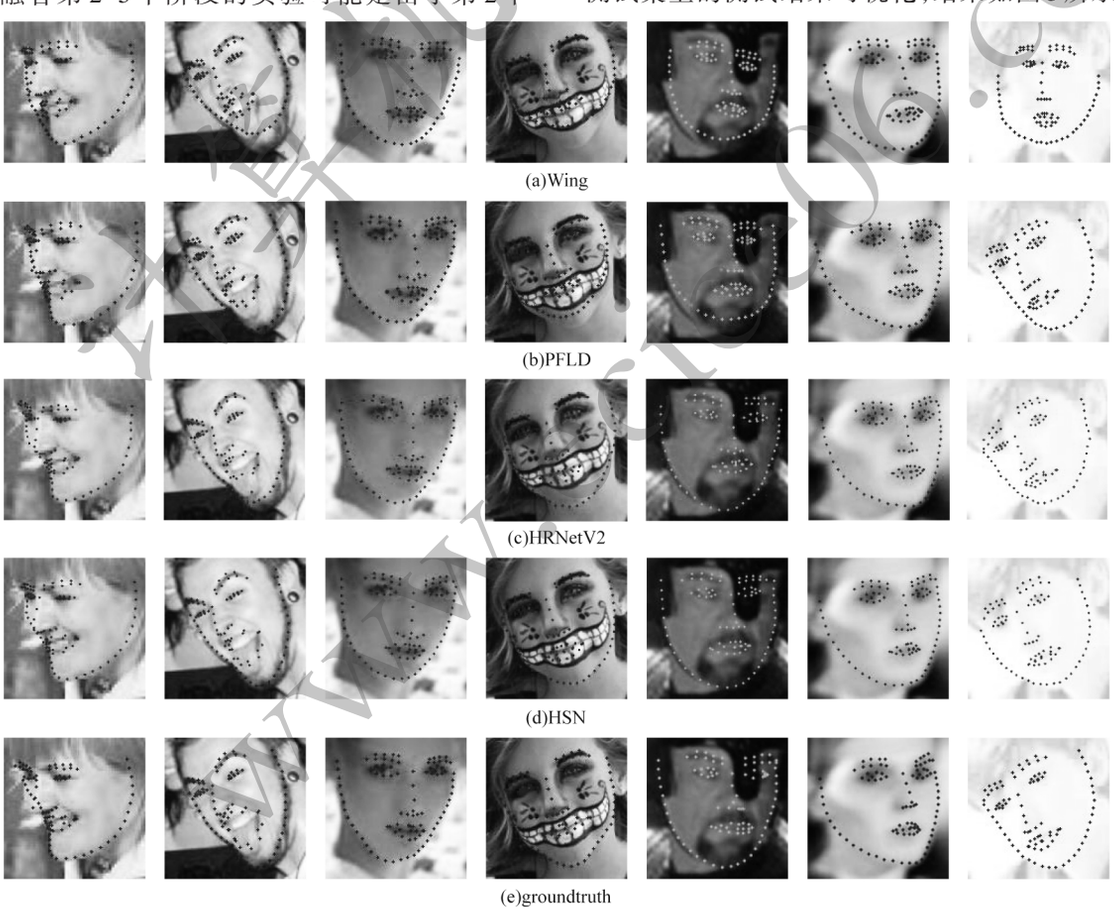

图5 不同算法的人脸特征点检测结果可视化对比

Fig.5 Visual results comparison of facial landmark detection using different algorithms

其中，Wing和PFLD是基于坐标回归方法，HRNetV2是基于热图回归方法。从图5可以看出，Wing、PFLD等坐标回归方法虽然保持整体形状的鲁棒性，但是在细节位置仍有较大偏差。Wing仅能预测大致形状，在大姿态角偏转等各种复杂情况下预测结果均不理想。PFLD对各个图中人脸整体外轮廓变化等全局变化以及眉毛走向、眯眼程度等局部变化反馈不到位。HRNetV2等基于热图回归方法

采用热图逐点预测保证了各点的精度，但丢失全局整体形状的鲁棒性，如图 5 中面部外轮廓均存在一定幅度的扭曲，第 2 列夸张表情示意图中将舌尖预测为嘴部特征点，第 7 列模糊 2 示意图中各器官扭曲等。因此，本文方法既保证整体形状的连续性，对大姿态偏转有良好的反馈性能，又能捕捉到各点局部细节，保证局部各点预测精度。

本文方法也存在一定的局限性。本文所提的多层次特征融合模块在优化特征点定位精度的同时，使得推理时间从原本的 13 ms 增加到 17 ms。该模块使用 Transformer 框架中的自注意力机制，当前各种芯片对该框架相关模型优化不足，使得在实际应用中所需推理时间可能进一步增加，这一点有待相关算子开发优化。此外，本文所提的多任务同时学习人脸特征点定位与姿态角的学习方式依赖数据集有角度上的多样性及姿态角数据，而大部分数据集缺乏该部分数据。虽然本文提出当数据集缺乏姿态角数据时的替代姿态计算算法，但此算法包含将平均人脸作为基准正脸和将人脸视为刚体的假设，在夸张表情等人脸数据与平均人脸相差较大时计算出的姿态角拟真值可能与实际值相差较大，影响模型拟合从而影响精度进一步提升，这一点有待包含姿态角数据的数据集扩充或姿态计算算法的进一步优化。

4 结束语

针对人脸特征点检测的基于坐标回归方法特征缺乏局部结构的细粒度表征能力、精度较低等问题，本文提出一种基于自注意力特征融合模块的网络算法。采用自注意力机制实现多层次特征融合，以实现不同阶段具有不同空间结构语义信息的特征层间交互，使得回归特征在全局形状具有鲁棒性的同时优化局部细粒度表征能力。提出一种多任务学习特征点检测定位及人脸姿态角估计的训练方式，提升算法对人脸整体姿态朝向的估计以提升特征点定位精度。下一步将优化姿态计算算法以适配更多数据集，研究网络的跨数据集泛化能力，提升网络的泛用性。

参考文献

[1] 于明，钟元想，王岩. 人脸微表情分析方法综述[J]. 计算机工程，2023，49(2)：1-14.
YU M，ZHONG Y X，WANG Y. A survey of facial micro-expression analysis methods[J]. Computer Engineering，2023，49(2)：1-14. (in Chinese)

[2] COOTES T F，EDWARDS G J，TAYLOR C J. Active appearance models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2001，23(6)：681-685.

[3] CRISTINACCE D，COOTES T F. Feature detection and tracking with constrained local models[C]//Proceedings of the British Machine Vision Conference. Edinburgh，UK：British Machine Vision Association，2006：1-10.

[4] NEWELL A，YANG K Y，DENG J. Stacked Hourglass networks for human pose estimation[C]//Proceedings of European Conference on Computer Vision. Berlin，Germany：Springer，2016：483-499.

[5] DOLLAR P，WELINDER P，PERONA P. Cascaded pose regression [C]//Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition. Washington D. C. ，USA：IEEE Press，2010：1-10.

[6] SUN Y，WANG X G，TANG X O. Deep convolutional network cascade for facial point detection[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ，USA：IEEE Press，2013：1-10.

[7] GUO X J，LI S Y，YU J K，et al. PFLD：a practical facial landmark detector[EB/OL]. [2023-01-10]. https://arxiv.org/abs/1902. 10859.

[8] FENG Z H，KITTLER J，AWAIS M，et al. Wing loss for robust facial landmark localisation with convolutional neural networks [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ，USA：IEEE Press，2018：2235-2245.

[9] ZHANG Z P，LUO P，LOY C C，et al. Facial landmark detection by deep multi-task learning[C]//Proceedings of European Conference on Computer Vision. Berlin，Germany：Springer，2014：94-108.

[10] YANG J，LIU Q S，ZHANG K H. Stacked Hourglass network for robust facial landmark localisation [C]//Proceedings of Conference on Computer Vision and Pattern Recognition Workshops. Washington D. C. ，USA：IEEE Press，2017：79-87.

[11] DENG J K，TRIGEORGIS G，ZHOU Y X，et al. Joint multi-view face alignment in the wild[J]. IEEE Transactions on Image Processing，2019，28(7)：3636-3648.

[12] WU W Y，QIAN C，YANG S，et al. Look at boundary：a boundary-aware face alignment algorithm[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ，USA：IEEE Press，2018：2129-2138.

[13] WANG X Y，BO L F，LI F X. Adaptive Wing loss for robust face alignment via heatmap regression [C]//Proceedings of International Conference on Computer Vision. Washington D. C. ，USA：IEEE Press，2019：6971-6981.

[14] LIN T Y，DOLLAR P，GIRSHICK R，et al. Feature pyramid networks for object detection [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ，USA：IEEE Press，2017：2117-2125.

[15] HARIHARAN B，ARBELáEZ P，GIRSHICK R，et al. Hypercolumns for object segmentation and fine-grained localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ，USA：IEEE Press，2015：447-456.

[16] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 3431-3440.

[17] XIE S N, TU Z W. Holistically-nested edge detection[C]// Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2015: 1395-1403.

[18] 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. New York, USA: ACM Press, 2017: 6000-6010.

[19] 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.

[20] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.

[21] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks[C]//Proceedings of the 14th European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 630-645.

[22] SAGONAS C, ANTONAKOS E, TZIMIROPOULOS G, et al. 300 faces in-the-wild challenge: database and results[J]. Image and Vision Computing, 2016, 47: 3-18.

[23] XU Z X, LI B H, YUAN Y, et al. AnchorFace: an anchor-based facial landmark detector across large poses[C]// Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2021: 3092-3100.

[24] ZHAO Y, LIU Y F, SHEN C H, et al. MobileFAN: Transferring deep hidden representation for face alignment[J]. Pattern Recognition, 2020, 100: 107114.

[25] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 5693-5703.

[26] YANG H L, YU J Y, CHENG P J, et al. LDDMM-Face: large deformation diffeomorphic metric learning for flexible and consistent face alignment[EB/OL]. [2023-01-10]. https://arxiv. org/abs/2108. 00690.

[27] ZHU C C, LI X Q, LI J D, et al. Improving robustness of facial landmark detection by defending against adversarial attacks[C]//Proceedings of International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2021: 11751-11760.

编辑 薛晋栋