

HCANet: 基于分层Transformer架构的微表情识别模型

杨澳男, 莫红, 赵世礼, 欧阳玉琦

(长沙理工大学电气与信息工程学院, 湖南 长沙 410114)

摘要: 面部微表情通过短暂肌肉运动展示了人类的真实情感, 具有微妙且不自觉的特点。为了探讨面部标记点固有的空间关联, 提高微表情识别准确率, 采用分层连续注意力网络 (hierarchical continuous attention network, HCANet) 以有效利用自注意力机制捕捉序列中标记点间的关系, 通过对起始帧与峰值帧的光流差异进行建模来避免直接从完整视频帧提取特征时忽视局部细节, 减少了对身份信息的干扰。HCANet主要由Transformer层和聚合层组成, 首先将人脸分为4个区域, 在Transformer层中, 引入连续注意力块 (continuous attention block, CAB) 专注于单个区域的局部细微肌肉运动用于提取本地时态特征; 其次通过跨层注意力传递机制, 在聚合层中专注于学习各区域之间的相互作用以提取全局语义面部特征; 最后在4个公开微表情数据集 (CASME II、CASME III、SMIC、SAMM) 上, 使用留一交叉验证与其他6个算法进行了对比验证。实验结果表明: HCANet在CASME III、SMIC、SAMM数据集上的分类准确率都有提升, 在复杂场景 (如低帧率、背景噪声) 中展现出更强的鲁棒性。

关键词: 微表情识别; 面部特征; 深度学习; 注意力机制

中图分类号: TP391.4

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202525

HCANet: a micro expression recognition model based on hierarchical Transformer architecture

YANG Aonan, MO Hong, ZHAO Shili, OUYANG Yuqi

School of Electrical & Information Engineering, Changsha University of Science & Technology, Changsha 410114, China

Abstract: Facial micro expressions are subtle, involuntary facial movements that reveal true emotions. To enhance recognition accuracy by exploiting spatial correlations among facial landmarks, a hierarchical Transformer architecture hierarchical continuous attention network (HCANet) was proposed to effectively leverage the self-attention mechanism for capturing relationships between landmarks in sequences. HCANet models optical flow differences between onset and apex frames to capture local details and reduce identity interference, thereby avoiding the oversight of local details when directly extracting features from full video frames. It consists of a Transformer layer for local temporal feature extraction and an aggregation layer for global facial feature learning. Initially, the face was divided into four regions. Within the Transformer layer, continuous attention block (CAB) was introduced to focus on the local, minute muscular movements within individual regions for extracting local temporal features. Subsequently, the aggregation layer concentrated on learning the inter-region interactions to extract global semantic facial features through a cross-layer attention mechanism. Finally, comparative validations were conducted using leave-one-out-cross-validation on four publicly available micro expression datasets (CASME II, CASME III, SMIC, SAMM) against six other algorithms. The results demonstrate that HCANet achieves improved classification accuracy on the CASME III, SMIC and SAMM datasets, and exhibits stronger robustness in complex scenarios (e.g., low frame rates, background noise).

Key words: micro expression recognition, facial feature, deep learning, attention mechanism

收稿日期: 2025-05-11; 修回日期: 2025-05-29

通信作者: 莫红, mohong198@163.com

0 引言

微表情指的是仅持续不到半秒（通常是 1/25 到 1/3 s）的细微肌肉运动，能揭示目标对象隐藏的真实情感，因此微表情识别技术被广泛应用于健康医疗、司法审讯等领域。近年来，基于计算机视觉的微表情识别方法得到了广泛的研究^[1]。虽然微表情数据集是在控制良好的实验室环境中收集的，但在 CASME III 数据集上，最新算法识别准确率甚至低于 50%，其识别准确率仍存在明显的提升空间^[2]。因此，研究人员仍在努力利用计算机视觉技术提高微表情识别的精确度和可靠性。

在微表情识别领域，一些研究人员采用局部二值模式（local binary pattern, LBP）方法提取面部特征，该方法基于纹理特征提取，在取得较好性能的同时，保持了较低的计算复杂度。此外，由于光流来源于连续帧之间的亮度差异，能够估计细微面部运动，也有相关学者探索了采用光流特征作为估计肌肉运动的输入的方法^[3]，包括双权重定向光流（bi-weighted oriented optical flow, BI-WOOF）^[4]、主方向平均光流（main directional mean optical flow, MDMO）^[5]、光流方向模糊直方图（fuzzy histogram of optical flow orientation, FHOFO）^[6]。许多著名的深度学习模型也被用于处理光流，甚至也有相关人员用强化学习方法^[7]来研究此类识别问题。

微表情识别包括两个主要阶段：面部特征提取和面部特征分类。传统分类方法依赖人工手动生成的特征^[8]，而深度学习方法则可以自动从面部图像中检测特征^[9-12]。为应对理解细微面部动作的难题以及训练数据可用量不足的情况，Liu 等^[3]提出利用领域自适应方法转移宏表情知识用于微表情识别。这包括放大微表情，并用更多生成的图像增强训练数据集，称为表情放大和缩小（expression magnification and reduction, EMR）方法。该方法在微表情识别挑战中表现优异。Liong 等^[13]使用浅层三流三维卷积神经网络（shallow triple stream three-dimensional convolutional neural network, STSTNet）来识别面部表情，其中光应变场、垂直光流场和水平光流场被送入这个浅层三流网络。此外，针对微表情识别，Xia 等^[14]提出采用循环卷积网络，该网络设计较浅，结合了循环神经网络（recurrent neural network, RNN）和卷积神经网络（convolutional

neural network, CNN）的优点，但是面部动作的细微性使得识别表情中特定面部肌肉和监测像素随时间的运动变得困难。为了解决这些问题，Zhou 等^[15]建议使用特征细化网络、特定表情特征学习网络和融合技术进行表情识别。此方法将自注意力机制应用于图像的全局特征上，可以提取微表情的显著特征。然而，图像特征之间的全局自注意力缺乏细粒度特征。

为了有效提取图像的全局自注意力细粒度特征，本文采用了一种新的分层连续注意力网络（hierarchical continuous attention network, HCANet），通过从起始帧和峰值帧的差异中提取肌肉运动模式特征。如图 1 所示，首先聚焦于面部的 4 个区域，即左眼区域、右眼区域、左唇区域和右唇区域^[16]，以消除实验室摄像机因潜在的灯光闪烁而接收到的背景噪声；其次，利用改进的 Transformer 层有效地捕获分层结构中的本地和全局交互，引入连续注意力块（continuous attention block, CAB）^[17]，专注于在身份信息较少的情况下对局部细微肌肉运动模式进行建模，Transformer 层中的低级自注意力旨在捕获局部区域内的细粒度特性，高级自注意力旨在捕获跨越全局区域的粗粒度特征；最后，为了促进同一级别不同块之间的交互，本文引入一个聚合块对不同块之间的特征进行融合。

该网络主要侧重于学习面部肌肉运动的位置和运动模式，且只对起始框架和顶点框架的差异进行建模，这反映了面部肌肉的运动，使模型受身份信息的影响较小。本文的贡献如下。

（1）本文引入了一种新的 Transformer 层自注意力机制，该机制有效地捕获了分层结构中的局部和全局交互，用于识别图像中的微表情。不同区域间的特征融合通过块聚合函数来实现。低层自注意力聚焦于局部区域的细粒度特征，高层自注意力聚焦于全局区域的粗粒度特征。

（2）本文改进了 Transformer 层的结构，通过增加连续注意力块提取肌肉运动模式特征，并利用聚合块将面部微表情的不同特征在网络末端进行融合。

（3）在 4 个数据集上开展实验，结果表明，HCANet 在检测数据集上的检测效果优于以往方法，这充分体现了该模型在微表情识别任务中的先进性与鲁棒性。

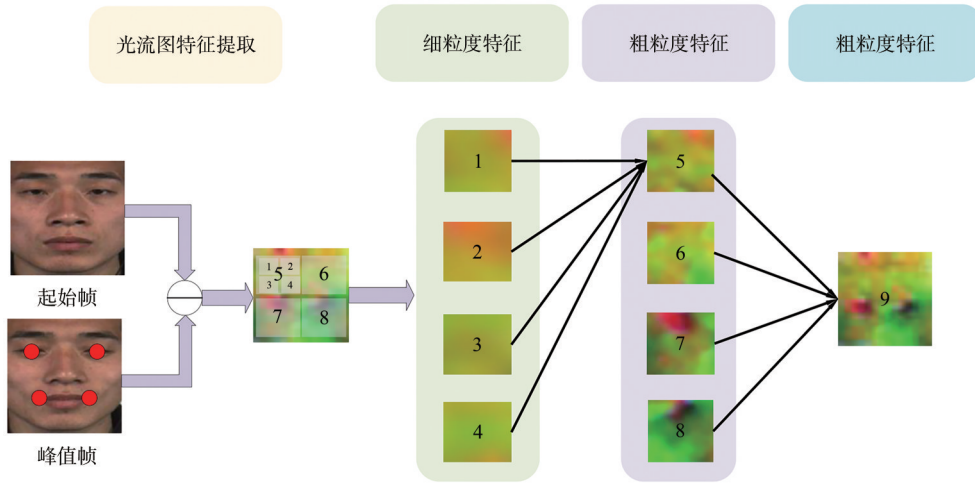


图1 HCANet框架

1 方法

1.1 光流图提取

为了获得面部微表情的光流图,需要确定起始帧和峰值帧索引。因为微表情数据集已经提供起始帧索引,所以还需要从视频序列中确定峰值帧索引。为此,采用了Liong等^[18]提出的分割处理感兴趣区域(divide and conquer-region of interest, D&C-ROI)方法,建立起始帧和后续帧之间的关系,以精准识别峰值帧索引,并为后续的微表情识别任务提供可靠的光流特征提取。

$$d = \frac{\sum_{i=1}^B h_{1i} \times h_{2i}}{\sqrt{\sum_{i=1}^B h_{1i}^2 \times \sum_{i=2}^B h_{2i}^2}} \quad (1)$$

其中, B 为直方图中矩形的个数, h_1 为第一帧, h_2 为除第一帧以外的其他帧。选取感兴趣区域(region of interest, ROI)差异率最高的作为代表面部肌肉变化最大的峰值帧。

此外,利用起始帧和峰值帧得到光流特征映射。光流图为:

$$V = (u(x, y), v(x, y)) \quad (2)$$

$$x = 1, 2, \dots, X, y = 1, 2, \dots, Y$$

其中, X 和 Y 分别表示帧的宽度 W 和高度 H , $u(x, y)$ 和 $v(x, y)$ 为光流特征映射 V 的水平和垂直分量, $V = [V_x, V_y]$, 其中 $V \in R^{W \times H \times 2}$ 。

本文使用光流场的一阶导数来计算光流场的变化,通常称为光应变。光应变提供了面部位移程度

的估计,这种光应变的计算能够捕捉和分析复杂的面部动态,有助于准确识别微表情。

$$V_z = \sqrt{\frac{\partial V_x^2}{\partial x} + \frac{\partial V_y^2}{\partial y} + \frac{1}{2} \left(\frac{\partial V_x^2}{\partial y} + \frac{\partial V_y^2}{\partial x} \right)} \quad (3)$$

其中, $\frac{\partial V_x^2}{\partial x}$ 、 $\frac{\partial V_y^2}{\partial y}$ 、 $\frac{\partial V_x^2}{\partial y}$ 和 $\frac{\partial V_y^2}{\partial x}$ 是 V 的一阶偏导数。

形成的三维光流特征映射表示为 $V_m = [V_x, V_y, V_z]$ 和 $V_m \in R^{W \times H \times 3}$ 。

为了从整个光流特征映射 V_m 中提取人脸光流特征映射,本文采用多任务级联卷积网络(multi-task cascaded convolutional network, MTCNN)^[19]从顶点图像中获取人脸地标坐标。随后从 V_m 中裁剪4个面部光流图,分别表示左眼光流图、左唇光流图、右眼光流图和右唇光流图。这4个光流图的大小都是 $\frac{W}{2} \times \frac{H}{2} \times 3$,是整个光流图大小的一半。在提取4个光流图特征之后,将它们组合起来,并将组合后的特征输入HCANet中进行微表情识别。整个过程如图2所示。

1.2 改进的Transformer层

与宏观表情相比,微表情的肌肉运动更小,局部活动区域更多,这使得传统的注意力模块难以准确地聚焦于细微的面部肌肉运动区域。为了解决这一问题,本文引入一个CAB,采用前一层的注意力图作为先验知识来生成当前层的注意力图。参考卷积注意力块模型(convolutional block attention module, CBAM)^[20]的计算方式,利用最大池化输出和平均池化输出来计算空间注意

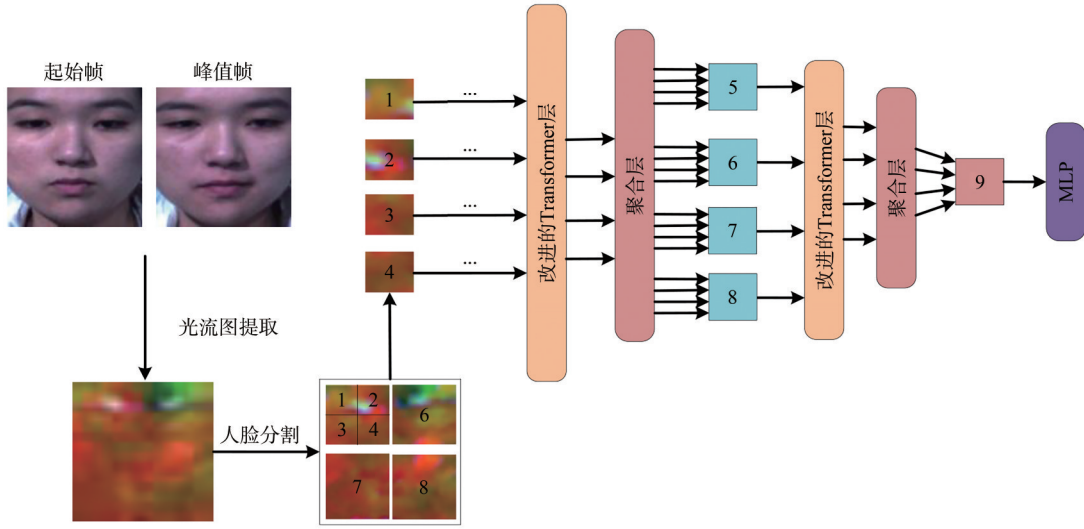


图2 HCANet: 分层的Transformer网络总体架构

图。如图3所示, 将前一层的注意力图作为先验知识来获得当前层的注意力, 并使用更小的卷积核(即 1×1)来获得更多的局部注意力。形式上, CAB可以定义为:

$$\text{Attn}^i = M^i(F_{\text{conv}}^i, \text{Attn}^{i-1}) = \sigma\left(f_{1 \times 1}^i\left(\left[P_{\text{Mc}}(F_{\text{conv}}^i); P_{\text{Ac}}(F_{\text{conv}}^i)\right]\right)\right) \otimes P_M(\text{Attn}^{i-1}) \quad (4)$$

$$F_{\text{conv}}^i = f_{1 \times 1}^i(f_{3 \times 3}^i(F^i)) + f_{1 \times 1}^i(F^i) \quad (5)$$

其中, M^i 为第 i 层CAB中期望关注肌肉运动区域的连续注意力模型, Attn^{i-1} 为第 $i-1$ 层的注意力图, $F_{\text{conv}}^i \in R^{2C \times H \times W}$ 表示第 i 层的前两层卷积提取的特征作为CAB的输入, σ 代表sigmoid函数, $f_{1 \times 1}^i$ 和 $f_{3 \times 3}^i$ 分别表示第 i 层卷积核大小为1和3的卷积操作, $P_{\text{Mc}}(F_{\text{conv}}^i) \in R^{1 \times H \times W}$ 和 $P_{\text{Ac}}(F_{\text{conv}}^i) \in R^{1 \times H \times W}$ 分别描述了跨通道的最大池化特征和平均池化特征, P_M 表示对第 $i-1$ 层的注意力图进行最大池化操作, 以匹配当前层注意力图的大小, $\otimes$ 表示将最后一层

的注意力图作为先验知识引入的元素积, $F^i \in R^{C \times 2H \times 2W}$ 表示第 i 个CAB的输入。

通过关联相邻层之间的注意机制, CAB可以连续不断且以较好的鲁棒性集中在具有细微运动的区域, 而不是集中在不同网络层对应的面部区域。用4个CAB构成Transformer层的主干, 学习细微肌肉运动模式特征 F_M , 大小为 $512 \times 14 \times 14$ 。由连续注意力模型和两个卷积层组成的CAB可以表示为:

$$\text{CAB}(F^i, \text{Attn}^{i-1}) = F_{\text{conv}}^i \otimes M^i(F_{\text{conv}}^i, \text{Attn}^{i-1}) \quad (6)$$

其中, CAB代表提出的连续注意力块, $\otimes$ 代表广播元素乘法, 即 $F_{\text{conv}}^i \in R^{2C \times H \times W}$ 的每个通道将乘以空间注意力映射 $M^i(F_{\text{conv}}^i, \text{Attn}^{i-1}) \in R^{1 \times H \times W}$ 关注感兴趣的区域。将峰值帧与起始帧之差作为主干的输入来学习运动模式特征。

当处理尺寸为 $H \times W \times 3$ 的输入光流图时, 每个图像块大小为 $P \times P$, 在这些光流图上进行线性投影和分割后, 每个块的特征维数为 $P \times P \times 3$, 之

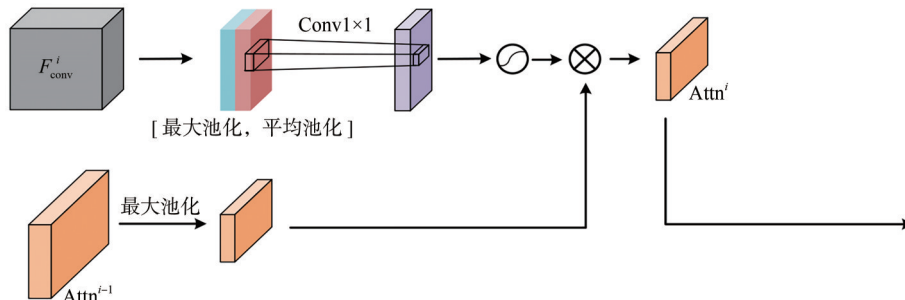


图3 CAB

后这些特征块被平整化处理, 模型的输入也就变成了 $X \in R^{b \times H_n \times n \times d}$, 其中, H_n 表示每一层的块数量, b 代表批次大小, d 是输入维数, n 是序列长度, 且 $H_n \times n = \frac{H \times W}{p^2}$ 。每一层 Transformer 由层归一化 (layer normalization, LN)、多头自注意力层 (multi-head self-attention, MSA) 和前馈全连接网络层 (feed-forward fully connected network, FFN) 组成, 其结构如图4所示。为了对空间信息进行编码, 将一个可训练的位置嵌入矢量加入所有的序列矢量中。这样可以确保空间关系和位置信息被有效地捕获并编码在特征表示中。

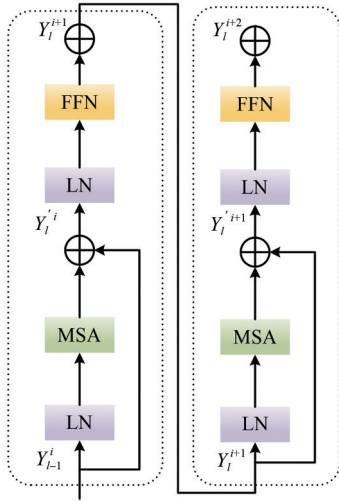


图4 Transformer层结构

$$\begin{aligned} Y_l^i &= Y_{l-1}^i + \text{MSA}\left(\text{LN}\left(Y_{l-1}^i\right)\right) \\ Y_l^{i+1} &= Y_l^i + \text{FFN}\left(\text{LN}\left(Y_l^i\right)\right) \\ Y_l^{i+1} &= Y_l^{i+1} + \text{MSA}\left(\text{LN}\left(Y_l^{i+1}\right)\right) \\ Y_l^{i+2} &= Y_l^{i+1} + \text{FFN}\left(\text{LN}\left(Y_l^{i+1}\right)\right) \end{aligned} \quad (7)$$

其中, $l=1, 2, \dots, L$, l 是每个层次 i 中的第 l 个块索引, L 是每一层中的块总数。

FFN 是包括 2 个线性层的全连接网络: $\max(0, xW_1 + b)W_2 + b$, 注意力机制应用在同一层的每个块 i 处, 在自注意力部分, 输入 $X \in R^{n \times d}$ 被分为 3 个部分, 即查询矩阵 Q 、键值矩阵 K 和值矩阵 V , 再对 Q 、 K 、 V 采用缩放点积:

$$\text{MSA}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \quad (8)$$

LN 层则按式 (9) 应用于每个块中:

$$\text{LN}(x) = \frac{x - \mu}{\delta} o\lambda + \beta \quad (9)$$

其中, μ 是特征平均值, δ 是特征标准偏差, o 是元素点, λ 和 β 是可学习的参数。

在引入 CAB 后, 改进后的 Transformer 层网络结构如图5所示。

1.3 聚合层

在 HCANet 模型中, 光流图的特征提取通过独立模块完成, 块聚合阶段则用来实现跨模块的信息

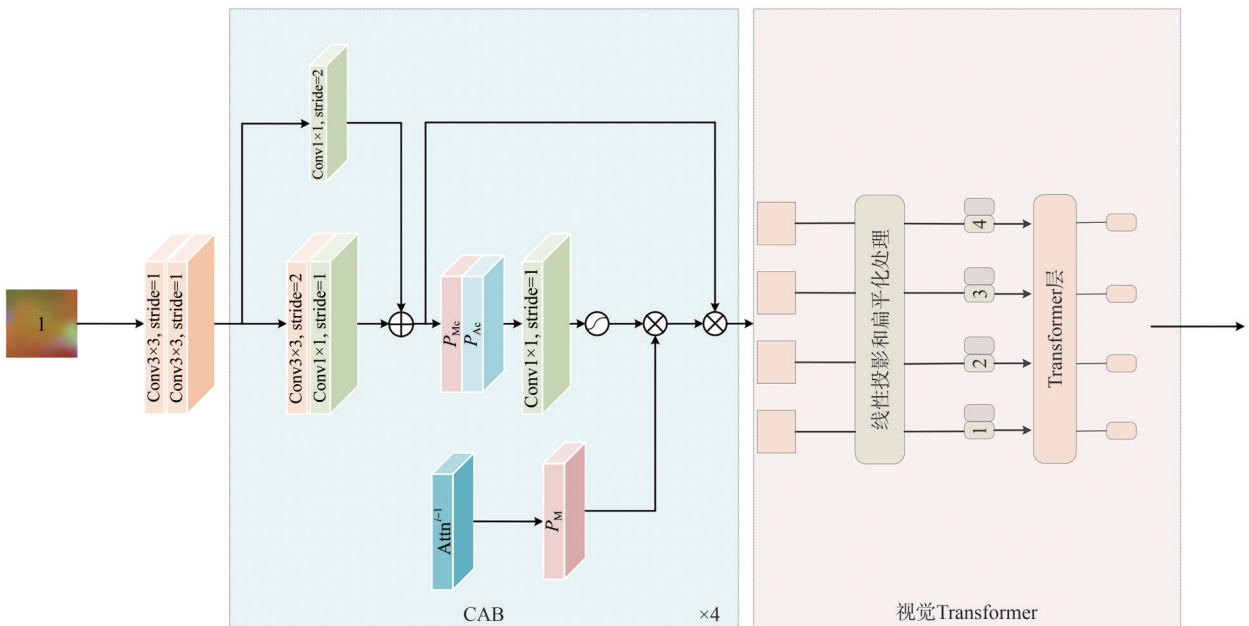


图5 改进后的Transformer层网络结构

交互。该过程以4个相邻块为操作单元，融合局部细节与全局语义特征。其中，低级聚合聚焦于左眼、右眼、左唇、右唇4个面部区域的内部特征交互，这个过程提取描述细微动态的细粒度信息，捕捉详细的面部动态；高级聚合则整合跨区域的全局特征，形成表征整体表情的粗粒度语义，促进4个面部区域之间的全局信息交换，整个过程如图6所示。初始输入的光流图尺寸为 $X_l \in R^{b \times h \times w \times d^l}$ 。经块聚合后输出的光流图大小为 $X_{l+1} \in R^{b \times 2h \times 2w \times d^{l+1}}$ ，其中 $d^{l+1} > d^l$ ，表明特征得到有效保留与增强。所有区域特征最终合并成完整面部特征图，记为 $X \in R^{b \times H \times W \times D}$ 。

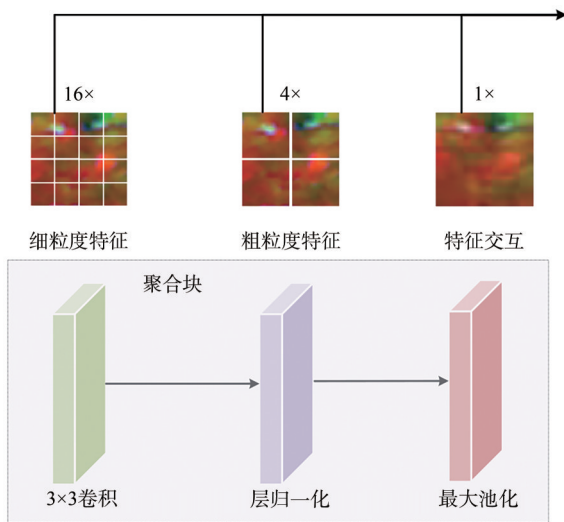


图6 聚合块结构

HCANet中的聚合块由卷积层、层归一化和最大池化构成。首先，初始阶段将人脸光流图划分为16个4×4特征块。首层3×3卷积负责对4个面部区域内特征进行局部融合，将块尺寸压缩至2×2；其次，次层3×3的卷积促进跨区域信息融合，进一步将区块压缩至1×1，并生成全面的光学图；最后，这些提取的光学图特征被送入多层感知器（multi-layer perceptron, MLP）层完成分类。这一分层策略通过多粒度特征捕捉和融合，有效地捕捉并融合

了不同层次的特征，对提升微表情识别性能具有积极作用。

1.4 损失函数

本文采用交叉熵损失函数来训练模型。交叉熵损失 L 的计算表示为：

$$L = \sum_i \left(-\omega_i \ln(p_t^i) \right)$$

$$p_t = p_c^y (1 - p_c)^{1-y} \quad (10)$$

其中， ω_i 为数据集样本的权重， y 为标签， $y_i=0,1$ 。

2 实验

2.1 数据集

为了验证HCANet的有效性，在4种流行的微表情数据集上进行实验，包括CASME II^[21]、SAMM^[22]、CASME III^[23]和SMIC^[24]。为了确保一致性和可比性，SAMM、SMIC和CASME II被合并成一个复合数据集，其中这3个数据集的相同情感标签被用于微表情识别任务。在这些数据集中，情绪类别划分如下：“积极”情绪类别包括“快乐”情绪类别，“消极”情绪类别包括“悲伤”“厌恶”“蔑视”“恐惧”和“愤怒”情绪类别，而“惊讶”情绪类别只包括“惊讶”情绪类别。这4个数据集的详细信息见表1，其中，消极、积极、惊讶分别代表各类情绪的样本数量，帧索引中的√、×代表此数据集是否提供相应的帧索引。

2.2 实验准备

本文实验采用多任务级联卷积网络（multi-task cascaded convolutional network, MTCNN）^[19]从顶点图像中提取人脸地标坐标，并根据这些地标点对图像进行裁剪。为了避免过拟合，从标记的起始帧和顶点帧周围的4帧中随机选择一帧作为起始帧和顶点帧进行训练，同时进行了水平翻转、随机裁剪和颜色抖动。利用Gunnar Farneback算法^[25]从这些图像中提取起始点和峰值时间点的光流。将光流图的3个元素——水平、垂直和光应变元素调整为28像素×28像素×3像素。在Transformer层中使用4个CAB

表1 4个数据集的详细信息

数据集	目标/人	样本/个	帧率/FPS	裁剪后图像分辨率	消极/个	积极/个	惊讶/个	起始帧索引	偏移帧索引	峰值帧索引
SAMM	32	133	200	28像素×28像素	92	26	15	√	√	√
CASME II	24	145	200	28像素×28像素	88	32	25	√	√	√
CASME III	16	164	100	28像素×28像素	70	51	43	√	√	×
SMIC	100	943	30	28像素×28像素	508	64	201	√	√	√

来构成主干, 并使用一个浅层视觉 Trasformer (vison Transformer, ViT) 来构建接收。基于人脸地标坐标, 提取了4个重要的人脸区域, 即左眼、左唇、右眼和右唇光流图。这4个光流图是整个光流图的一半大小, 即14像素×14像素×3像素。在提取4个光流图之后, HCANet接收并组合它们。这种全面的方法确保了HCANet能够准确有效地识别和分类微表情。

本文实验在 Windows10 操作系统上使用 PyTorch 和 Python 3.6 进行。将训练参数的学习率设置为 5×10^{-5} , 并使用最多 600 个训练轮次 (epoch) 进行训练。

2.3 实验结果

复合微表情数据集中的类分布是不平衡的, 不同的情绪有不同的频率。具体来说, 惊讶、积极和消极情绪的比例分别约为 1:1.3:3。为了解决这个问题, 本文采用了未加权 F1 分数 (unweighted F1-score, UF1) 和未加权平均召回率 (unweighted average recall, UAR) 作为模型的性能指标。

(1) UF1 是一种常用于评估具有不平衡分布的多类任务性能指标。要计算 UF1, 先要统计每个类别的假阳性 (false positive, FP)、真阳性 (true positive, TP) 和假阴性 (false negative, FN) 样本数量, 然后由式 (11) 得出:

$$F1_i = \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i}$$

$$UF1 = \sum_{i=1}^C \frac{F1_i}{C} \quad (11)$$

其中, i 表示第 i 类样本, C 是样本总类别数, $F1$ 表示 $F1$ 分数。

(2) UAR 在评估不平衡类比率的模型时十分有效。要计算 UAR, 先要统计每个类别的 TP 和样本总数 N , 然后由式 (12) 得出:

$$Recall_i = \frac{TP_i}{N_i}$$

$$UAR = \frac{1}{C} \sum_{i=1}^C Recall_i \quad (12)$$

其中, Recall 表示召回率。

为了验证分层结构的有效性, 本文设置了一个对未经分割的光流图直接进行识别的单层网络架构, 同时保留相同参数量。实验结果见表 2, 相较于单层网络架构, 分层结构在 CASME II、SAMM、SMIC 数据集上 UAR/UF1 性能指标分别提升了 0.0221/0.0361、0.0375/0.0238、0.0255/0.0176。

表 2 分层结构与单层结构比较

方法	CASME II		SAMM		SMIC	
	UAR	UF1	UAR	UF1	UAR	UF1
单层结构	0.8002	0.7231	0.7063	0.6254	0.8132	0.8042
分层结构	0.8223	0.7592	0.7438	0.6492	0.8387	0.8218

为了证明将前一层的注意力图作为先验知识引入当前的微表情识别 (micro expression recognition, MER) 任务注意力图上的有效性, 将提出的连续注意模块与传统的独立注意模块 (即 CBAM 的空间注意模块) 进行了比较, 后者在每一层独立生成空间注意力图。实验结果见表 3, CAB 在 CASME II、SAMM、SMIC 数据集上的性能都明显优于独立关注模块, 其 UAR/UF1 性能指标分别提升了 0.0505/0.0587、0.0845/0.0293、0.0651/0.0924。

表 3 连续注意力与独立注意力比较

方法	CASME II		SAMM		SMIC	
	UAR	UF1	UAR	UF1	UAR	UF1
独立注意力	0.8368	0.8126	0.7429	0.6876	0.7998	0.7568
连续注意力	0.8873	0.8713	0.8274	0.7169	0.8649	0.8492

初始实验设置底层块大小为 7×7 , 隐藏维度为 256, Transformer 中设置 2 个多头, 每个层中 2 个 Transformer 编码器 (encoder) 层。针对每个消融实验, 修改一个特定的参数, 同时保持其他设置不变。

表 4 显示了不同超参数设置对微表情识别准确率的影响。 N_L 和 N_H 分别代表 Transformer encoder 层数和多头数量, 可以观察到较小的 N_L 和 N_H 倾向于获得更好的结果。这可能是因为较大的模型更容易过度拟合。因此, 默认设置 Transformer encoder 层数 N_L 和自注意多头数 N_H 分别为 2 和 4。

表 4 Transformer 中层数和多头数量对实验结果的影响

组别	N_L	N_H	SAMM	CASME II	CASME III	SMIC
1	2	2	78.56%	84.25%	59.23%	79.21%
2	2	4	81.03%	87.68%	61.56%	83.36%
3	2	8	77.46%	86.55%	63.43%	81.02%
4	3	2	81.88%	84.03%	62.65%	80.95%
5	3	4	79.26%	84.54%	60.21%	79.93%
6	3	8	74.21%	82.58%	58.91%	76.44%

本文还通过 4 组不同数据比较了 HCANet 在微表情识别任务中的综合性能。从表 5 可以看出, 在与其他模型的对比中, HCANet 在大部分评价指标

上有所提升,相较于目前最先进的(state-of-the-art, SOTA)模型微型双向编码器网络(micro bidirectional encoder representations from transformer, μ -BERT)^[26],在SMIC数据集上UF1和UAR指标分别提升0.012和0.073;在样本规模更大的CASME II数据集上UF1和UAR指标取得0.0259和0.0298的性能增益;值得注意的是,在最新的CASME III大规模数据集上,UF1和UAR指标分别实现了0.2778和0.2693的显著提升,这一结果验证了HCANet在复杂光照、多视角场景下的鲁棒性特征学习能力。即便在样本异质性较强的SAMM数据集上,HCANet仍保持0.0110的UF1指标提升,说明模型具备良好跨数据库泛化性能。

3 分析

图7(a)为HCANet混淆矩阵,显示了在完整的复合数据集中每个情感类别的准确性。值得注意的是,在SMIC、SAMM和CASME II数据集中,HCANet的消极、积极和惊讶类别的准确率分别为0.95、0.70和0.81。消极类别的高准确率可归因于3个数据集中该类的可用训练样本数量较多。然而,在区分SMIC和SAMM数据集中的惊讶和积极情绪类别方面仍存在挑战。这些数据集中的训练样本数量有限,可能会导致这两个类别之间偶尔出现错误分类。此外,CASME II数据集提供了更精确的峰值帧,产生了更准确的光流图,从而更好地捕捉了面部肌肉运动。因此,CASME II数据集中的消极、积极和惊讶类别达到了相对较高的精度。此外,SMIC数据集使用较低的帧率来捕获图像,引入了背景噪声,如闪烁的灯光、阴影和照明变化,从而影响了SMIC数据集中积极情绪和惊讶情绪分类的准确性,导致准确率分别为65%和77%。本文实验选取的不同数据集样本数量存在差距,因此

在某些数据集上的提升效果不够明显,而小样本数据集在特征提取过程中引入迁移学习^[27]通常会使视觉特征的分类效果更好,因此有望进一步提升识别准确率。

4 结束语

本文提出了一种新的基于分层Transformer架构的微表情识别模型。与以前的深度学习方法相比,HCANet的一个关键优势是能够模拟长期时间依赖性,这对于准确识别经常在短时间内发生的微表情至关重要。该模型的层次结构允许在多个时间尺度上提取特征,这可以通过捕获表达式中的短期和长期动态来提高识别精度。此外,基于CAB的主干专注于从起始帧和顶点帧的差异中学习运动模式特征,有效抑制了背景噪声和与主体无关的面部动作干扰。

尽管如此,HCANet在工程落地中仍面临两大挑战:其一,模型对高质量标注数据的依赖程度较高,而微表情数据的采集需依赖专业设备与心理诱导实验,成本高昂;其二,注意力机制在非受控环境(如侧脸、遮挡场景)下的识别精度仍有待提升。未来研究可从技术改进、数据范式创新及多学科交叉3个维度进一步推动微表情识别的发展。首先,在模型结构和训练范式优化上,探索知识蒸馏或神经架构搜索技术,在保证精度的前提下压缩模型参数量,使其适配移动端或边缘计算设备,满足安防监控、远程医疗等场景的实时分析需求;其次,在数据获取与增强策略创新上,开发基于生成对抗网络(generative adversarial network, GAN)的微表情合成引擎,通过控制动作单元(action unit, AU)编码与持续时间参数,生成高保真、多样化的训练样本,需特别关注合成数据与真实数据的域差异消减问题;最后,在应用场景拓展与跨学

表5 HCANet与其他模型性能对比

模型	SAMM		CASME II		CASME III		SMIC	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
AlexNet	0.6104	0.6642	0.7994	0.8312	—	—	0.6201	0.6373
VCG16	0.4870	0.4793	0.8166	0.8202	—	—	0.5800	0.5964
OFF-ApexNet	0.5409	0.5392	0.8764	0.8681	—	—	0.7196	0.7096
STSTNet	0.6588	0.6810	0.8382	0.8686	0.3795	0.3792	0.6801	0.7013
RCN	0.7601	0.6715	0.8512	0.8123	0.3928	0.3893	0.6326	0.6441
μ -BERT	0.8386	0.8475	0.9034	0.8914	0.3264	0.3254	0.8550	0.8384
HCANet	0.8496	0.8337	0.9293	0.9212	0.6042	0.5947	0.8562	0.8457

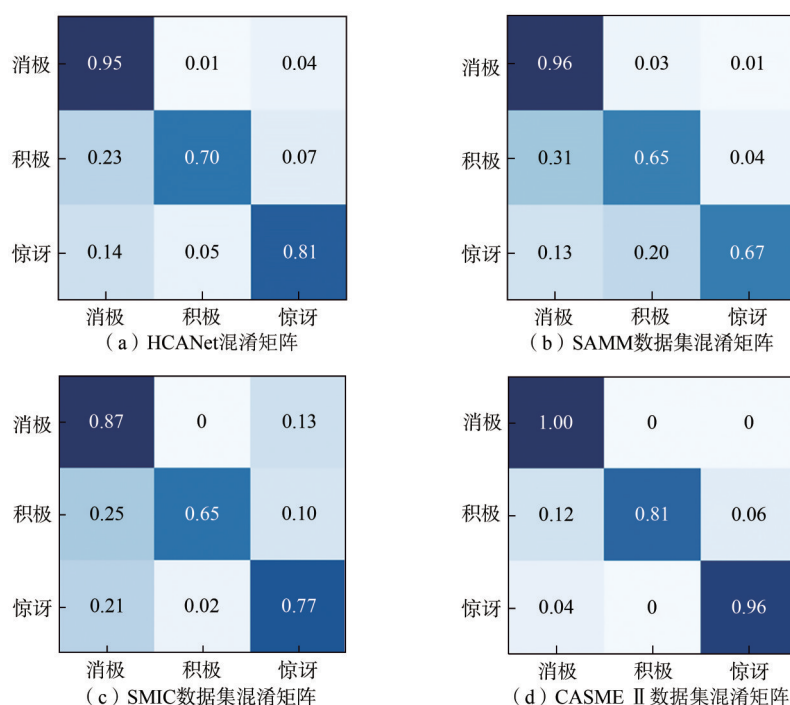


图7 HCANet在不同数据集上的混淆矩阵

科研究方面, 开发对抗性训练框架, 联合优化光照归一化、头部姿态估计与微表情识别模块, 提升复杂环境下的模型泛化能力。

通过以上方向的持续探索, 微表情识别技术有望从实验室走向实际应用, 最终在情感计算、智能医疗、公共安全等领域形成闭环落地价值。

参考文献:

- [1] WANG Y, SUN Y X, HUANG Y W, et al. FERV39k: a large-scale multi-scene dataset for facial expression recognition in videos[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 20890-20899.
- [2] GAN Y S, LIONG S T, YAU W C, et al. OFF-ApexNet on micro-expression recognition system[J]. Signal Processing: Image Communication, 2019, 74: 129-139.
- [3] LIU Y C, DU H M, ZHENG L, et al. A neural micro-expression recognizer[C]//Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Piscataway: IEEE Press, 2019: 1-4.
- [4] LIONG S T, SEE J, WONG K, et al. Less is more: Micro-expression recognition from video using apex frame[J]. Signal Processing: Image Communication, 2018, 62: 82-92.
- [5] LIU Y J, ZHANG J K, YAN W J, et al. A main directional mean optical flow feature for spontaneous micro-expression recognition[J]. IEEE Transactions on Affective Computing, 2016, 7(4): 299-310.
- [6] HAPPY S L, ROURAY A. Fuzzy histogram of optical flow orientations for micro-expression recognition[J]. IEEE Transactions on Affective Computing, 2019, 10(3): 394-406.
- [7] 王飞跃, 曹东璞, 魏庆来. 强化学习: 迈向知行合一的智能机制与算法[J]. 智能科学与技术学报, 2020, 2(2): 101-106.
- [8] WANG F Y, CAO D P, WEI Q L. Reinforcement learning: toward action-knowledge merged intelligent mechanisms and algorithms[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(2): 101-106.
- [9] ZHAO G Y, PIETIKÄINEN M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 915-928.
- [10] LEI L, CHEN T, LI S G, et al. Micro-expression recognition based on facial graph representation learning and facial action unit fusion[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2021: 1571-1580.
- [11] ZHANG K H, HUANG Y Z, WU H, et al. Facial smile detection based on deep learning features[C]//Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Piscataway: IEEE Press, 2015: 534-538.
- [12] ZHANG K H, HUANG Y Z, DU Y, et al. Facial expression recognition based on deep evolutionary spatial-temporal networks[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4193-4203.
- [13] NIU W J, ZHANG K H, LI D X, et al. Four-player GroupGAN for weak expression recognition via latent expression magnification[J]. Knowledge-Based Systems, 2022, 251: 109304.
- [14] LIONG S T, GAN Y S, SEE J, et al. Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition[C]//Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Piscataway: IEEE Press, 2019: 1-5.
- [15] XIA Z Q, PENG W, KHOR H Q, et al. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 8590-

- 8605.
- [15] ZHOU L, MAO Q R, HUANG X H, et al. Feature refinement: an expression-specific feature learning and fusion method for micro-expression recognition[J]. Pattern Recognition, 2022, 122: 108275.
- [16] WANG Z F, ZHANG K H, LUO W H, et al. HTNet for micro-expression recognition[J]. Neurocomputing, 2024, 602: 128196.
- [17] LI H T, SUI M Z, ZHU Z Q, et al. MMNet: Muscle motion-guided network for micro-expression recognition[EB]. arXiv preprint, 2022: arXiv: 2201.05297.
- [18] LIONG S T, SEE J, WONG K, et al. Automatic apex frame spotting in micro-expression database[C]//Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Piscataway: IEEE Press, 2015: 665-669.
- [19] JOSE E, M G, HARIDAS M T P, et al. Face recognition based surveillance system using FaceNet and MTCNN on jetson TX2[C]//Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). Piscataway: IEEE Press, 2019: 608-613.
- [20] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). Berlin: Springer, 2018: 3-19.
- [21] YAN W J, LI X B, WANG S J, et al. CASME II: an improved spontaneous micro-expression database and the baseline evaluation[J]. PLoS One, 2014, 9(1): e86041.
- [22] DAVISON A K, LANSLEY C, COSTEN N, et al. SAMM: a spontaneous micro-facial movement dataset[J]. IEEE Transactions on Affective Computing, 2018, 9(1): 116-129.
- [23] LI J T, DONG Z Z, LU S Y, et al. CAS(ME)₃: a third generation facial spontaneous micro-expression database with depth information and high ecological validity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 2782-2800.
- [24] LI X B, PFISTER T, HUANG X H, et al. A spontaneous micro-expression database: inducement, collection and baseline[C]//Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Piscataway: IEEE Press, 2013: 1-6.
- [25] CHANG C W, ZHONG Z Q, LIOU J J. A FPGA implementation of farneback optical flow by high-level synthesis[C]//Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. New York: ACM Press, 2019: 309-309.
- [26] NGUYEN X B, DUONG C N, LI X, et al. Micron-BERT: BERT-based facial micro-expression recognition[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Piscataway: IEEE Press, 2023: 1482-1492.

- [27] 刘栋军, 王宇涵, 凌文芬, 等. 基于脑机协同智能的情绪识别[J]. 智能科学与技术学报, 2021, 3(1): 65-75.

LIU D J, WANG Y H, LING W F, et al. Emotion recognition based on brain and machine collaborative intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2021, 3(1): 65-75.

[作者简介]



杨澳男 (1999-), 男, 长沙理工大学电气与信息工程学院硕士生, 主要研究方向为数字图像与机器视觉。



莫红 (1972-), 女, 长沙理工大学电气与信息工程学院教授, 主要研究方向为智慧医疗、模糊AI和复杂系统管理与控制。



赵世礼 (2000-), 男, 长沙理工大学电气与信息工程学院硕士生, 主要研究方向为数字图像与机器视觉。



欧阳玉琦 (2000-), 男, 长沙理工大学电气与信息工程学院硕士生, 主要研究方向为数字图像与机器视觉。