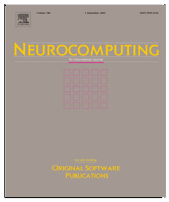
Neurocomputing 602 (2024) 128196

可在ScienceDirect上找到的目录列表

神经计算

期刊主页：www.elsevier.com/locate/neucom



用于微表情识别的HTNet

*Zhifeng Wanga，*Kaihao Zhanga，Wenhan Luob，Ramesh Sankaranarayanaa澳大利亚国立大学工程与计算机科学学院，堪培拉，澳大利亚*

b中国广州中山大学

ARTICLE INFO

由王思杰传达

关键词：
层次变换器微表达识别深度学习
面部肌肉运动局部自关注

ABSTRACT

面部表情与面部肌肉收缩有关，不同的肌肉运动对应着不同的情绪状态。对于微表情识别，肌肉运动通常较为微妙，这对接目前面部情感识别算法的性能产生了负面影响。大多数现有方法使用自注意力机制来捕捉序列中标记之间的关系，但它们没有考虑到面部特征点之间的固有空间关系。这可能导致在微表情识别任务上的表现不佳。因此，学习识别面部肌肉运动是微表情识别领域的一个关键挑战。本文提出了一种层次变换网络（HTNet），用于识别面部肌肉运动的关键区域。HTNet包含两个主要组件：一个利用局部时间特征的变换层和一个提取局部和全局语义面部特征的聚合层。具体来说，HTNet将脸部分为四个不同的区域：左唇区、左眼区、右眼区和右唇区。变换层用于通过每个区域的局部自注意力机制来表示局部细微的肌肉运动。聚合层则用于学习眼睛区域与嘴唇区域之间的相互作用。在四个公开可用的微表情数据集上的实验表明，所提出的方法比以往的方法有显著提升。代码和模型可在以下网址获取：https://github.com/wangzhifengharrison/HTNet。

1. 介绍

微表情指的是持续时间仅为大约1/25-1/5秒的微妙肌肉运动。近年来，大量研究致力于利用计算机视觉方法分析微表情[1,2]。然而，识别微表情的准确性仍有待提高。尽管微表情数据集是在严格控制的实验室环境中收集的，但目前的结果仍不尽如人意[3]。由于微表情通常伴随细微的肌肉运动，这使得人类和计算机都难以检测到，因此利用计算机视觉进行这项任务仍然具有挑战性。另一方面，正常宏观表情的识别已经达到了很高的准确率（超过95%）[4]。这种性能上的巨大差异可以归因于微表情难以分析的事实，因为它们具有短暂性和细微特征。因此，研究人员仍在努力提高使用计算机视觉技术识别微表情的准确性和可靠性。

在微表情识别领域，一些研究人员提出了采用局部二进制模式（LBP）方法提取面部特征。LBP技术通过其基于纹理的特征提取方法展示了令人称赞的区分能力，同时保持了较低的计算复杂度[5]。此外，其他学者探索了使用光流特征作为输入来估计肌肉运动[5]。光流是从连续帧之间的亮度差异中得出的，能够估计细微的面部动作。已经研究了几种基于光流的方法，包括BiWOOF [6]，MDMO [7]，FHOFO [8]，光学应变权重和光学应变特征[9]。此外，诸如VGG16 [10]、GoogleNet [11]、AlexNet [12]和OFF-Apex [3]等显著的深度学习模型已被用于处理TV-L1光流。这种光流是从图像序列中选定的顶点和起始帧中提取的。顶点帧捕捉了关于微表情最显著的信息，对于准确识别表情至关重要。

在面部表情识别领域，由于面部动作的微小和微妙性质，分析微表情具有挑战性。这使得识别表情所涉及的具体面部肌肉和监测像素随时间的运动变得困难。

*通讯作者。
电子邮箱：zhifeng.wang@anu.edu.au（王志），super.khzhang@gmail.com（张科），whluo.china@gmail.com（罗伟），ramesh.sankaranarayana@anu.edu.au (R. Sankaranarayana).

https://doi.org/10.1016/j.neucom.2024.128196
接收日期：2023年7月27日；修订后接收日期：2024年4月16日；接受日期：2024年7月14日
2024年7月23日在线发布

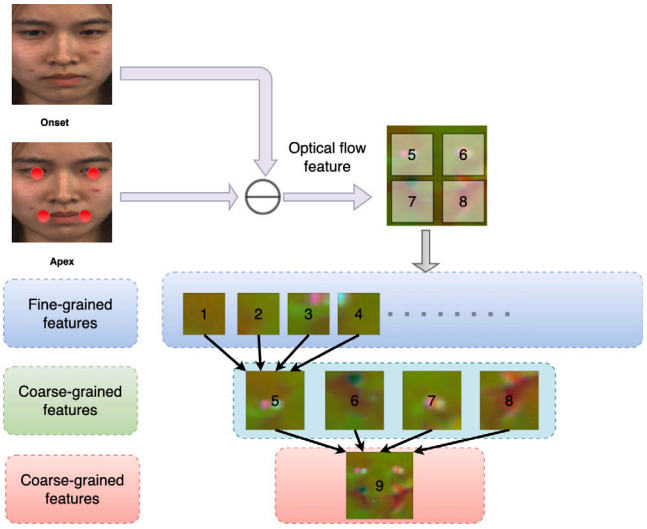


图1.提出的HTNet范式。第一步，我们使用起始帧和顶点帧获取光流特征。第二步，通过面部关键点获取中间级别的粗粒度特征。这些中间级别的粗粒度特征将均匀分为四部分，即细粒度特征。我们使用聚合块来结合不同图像尺度下的局部细粒度和全局粗粒度交互。通过这一新机制，同一层级块中的每个像素以细粒度处理，而上层块中的像素则以粗粒度处理。这使得我们的模型能够有效捕捉短距离和长距离的视觉依赖关系。

针对这些问题，周等人[13]建议使用特征精炼网络、特定表情特征学习网络以及融合技术进行表情识别。通过在全球图像特征上应用自注意力机制，他们的方法能够提取微表情识别的显著特征。然而，图像特征之间的全局自注意力机制在多尺度上缺乏精细特征。此外，它们没有考虑面部特征点之间的固有空间关系。这可能导致在微表情识别任务中表现不佳。为了解决这些问题，我们的目标是在每个层次结构的局部块中保持自注意力机制，以捕捉图1中的细粒度特征。我们使用聚合块来结合不同图像尺度下的局部细粒度和全局粗粒度交互。通过这种新机制，同一层级块中的每个像素都以细粒度处理，而上级块中的像素则以粗粒度处理。这使得我们的模型能够有效捕捉短距离和长距离的视觉依赖关系。在这项研究中，我们提出了一种利用Transformer层的新自注意力方法，以有效捕捉层次结构内的局部和全局交互。Transformer层中的低级自注意力旨在捕捉局部区域内的细粒度特征。另一方面，这些层中的高级自注意力则设计用于捕捉跨越全局区域的粗粒度特征。为了促进同一层级不同模块之间的交互，我们提出了一种聚合模块。我们提出的HTNet方法的整体架构如图2所示。本研究的贡献可以总结如下：

(a) 我们介绍了一种新颖的自注意力机制，该机制通过Transformer层有效地捕捉层次结构中的局部和全局交互，以识别图像中的微表情。这是通过利用所提出的块聚合函数实现的。低级自注意力专注于捕捉局部区域的细粒度特征，而高级自注意力则针对全局区域的粗粒度特征。

(b) 我们的网络专门关注四个面部区域——左眼区域、左唇区域、右眼区域和右唇区域——而不是考虑整个面部区域。这种方法有助于以减轻背景噪声对面部区域边缘的影响，这些噪声可能被实验室摄像头捕捉到。通过使用Transformer层，我们可以专注于通过局部自注意力模型来模拟每个区域内的小规模、细微的肌肉运动。此外，聚合层学习眼睛区域和嘴唇区域之间的相互作用。我们进行实验来研究不同的块大小如何影响微表情识别的准确性。

(c) 通过在四个可用数据集上进行的实验，我们证明了我们的方法明显优于以前的方法。这突出了我们的模型在微表情识别任务中的有效性和优越性。

2. 相关工作

2.1. 手工制作的功能

在接下来的小节中，我们将讨论两种主要的手工特征：基于外观的特征和基于几何的特征。

2.1.1. 基于外观的特征

在传统的表达识别技术中，通常利用基于外观的特征[14－16]。一种常见的模式是从三个正交平面（LBPTOP）提取局部二值图案[17]。在一些LBP-TOP研究中，LBP-TOP被转移到与张量无关的RGB空间，这增强了鲁棒性[18]。为了降低计算复杂度，LBP-SIP有效地减少了LBP-TOP模式中的冗余，提供了一种轻量级的表示[19]。TICS [20]引入了一种新的颜色模型，称为张量独立色彩空间（TICS），以提高其识别真实情感的能力。他们的方法将微表情的彩色视频片段视为一个四维数组，通过将常规的RGB颜色维度转换为TICS，从而获得更独立的颜色组件，进而实现对这些快速表情的更准确识别。

2.1.2. 基于几何的特征

基于几何特征通常分为两类：基于光流的和基于纹理变化的特征。这些基于几何的特征可以识别运动变形。李等人[21]探索了一种用于定位面部标志点并划分面部区域的兴趣区的深度学习技术。由于面部微表情是由面部肌肉收缩产生的，评估这些收缩的方向对于识别情绪至关重要。利用这一技术，他们将面部划分为与不同肌肉动作模式相关的关注区域。此外，面部活动可以通过光流充分反映。刘等人。[7]提出使用强大的MDMO特征提取网络来识别微表情。他们计算视频序列中每张图像的光流，并将面部区域划分为多个引人入胜的部分。然后从每张图像中计算平均光流特征，并对这些平均光流特征应用SVM分类器以识别情绪。他们的方法有效地考虑了地理位置和区域统计运动信息，证明既简单又高效。Liong等人[6]不采用LBP直方图，而是建议加权LBP直方图并平均光流特征值以识别面部情绪，即Bi-WOOF。Wang等人[20]提出利用张量独立颜色空间中的动态纹理进行微表情识别。他们通过将彩色视频片段转换为四维数组来分析空间、时间和独立颜色成分，从而提高微表情识别的准确性，优于传统的RGB颜色空间。Lu等人[22]呈现了一种基于德劳内法的时间

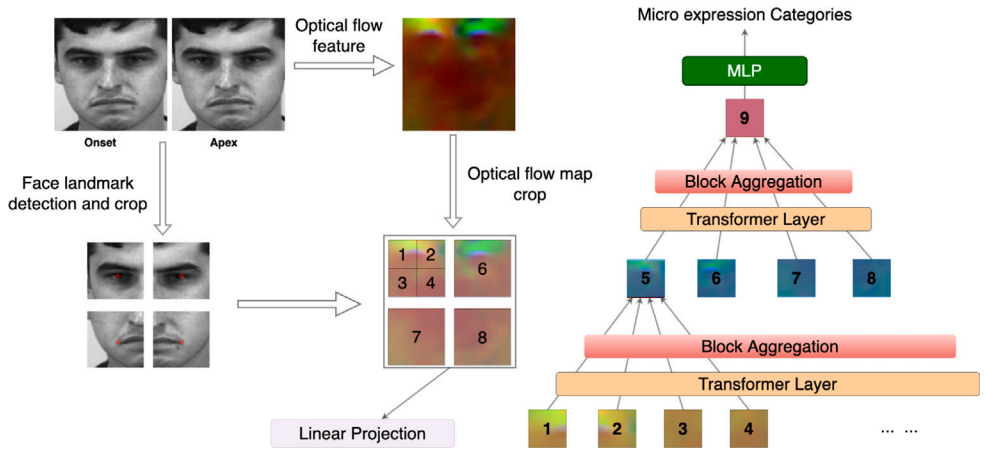


图2. HTNet：用于微表情识别的层次变换器网络的整体架构。低级自注意力机制在变换层中捕捉局部区域的细粒度特征。高级自注意力机制在变换层中捕捉全局区域的粗粒度特征。提出了一种聚合块，以在同一层级的不同块之间创建交互。

编码模型（DTCM）通过归一化面部图像序列并抑制无关特征，有效编码与肌肉活动相对应的纹理变化，从而提高识别率，这一点已在主要数据集上的广泛实验中得到验证。然而，手工设计的特征提取通常需要专家设计的提取器，这一过程往往需要专业知识。

2.2. 深度学习方法

近年来，一系列深度学习技术已经出现，用于提取表情识别[23－26]中的面部特征。[27]提出应用循环卷积网络来建立面部位置信息与记录不同区域面部肌肉收缩之间的联系。该模型结合了多个循环卷积层和一个分类层，以捕捉用于面部情感识别的视觉特征。Gan等人[3]提出了一种自动方法来定位顶点帧，并使用光流作为其OFF-ApexNet的输入。利用老化卷积神经网络，他们的网络从光流中提取新的特征描述符。另一方面，一些研究人员探索了浅层卷积神经网络的应用，这些网络能够从运动估计的三个组成部分——水平、垂直光流场和光学应变中有效推导出高层次的视觉属性，从而推断情绪状态[28]。为了应对复合数据库领域偏移的问题，Xia等人[29]开发了一个RCN模型，研究较小模型如何影响微表情的识别。在RCN中，他们设计了三个无参数模块，包括注意力单元、捷径连接和宽扩展，以防止可学习参数数量的增加。然而，卷积这一用于图像分析的过程总是基于固定的窗口大小，这意味着它可能无法捕捉到相距较远的关系或模式。为了解决这个问题，Kumar等人[30]提出了一种新颖的双流图注意力网络，利用面部特征点与光流补丁之间的关系来检测和分类细微且短暂的面部微表情。此外，Zhang等人[31]引入了一种新的架构，使用变压器而不是传统的卷积网络来识别微表情。该架构包括一个空间编码器以学习空间模式，一个时间聚合器以分析时间，并有一个分类头。但是，他们的方法忽略了面部特征点之间的空间关系，这可能导致微表情识别效果不佳。同时，Lei等人。[23]引入了Transformer作为编码器来建模人脸节点和边之间的连接，基于人脸特征的边和节点构建人脸图。然而，它们不考虑在层次结构中提取特征。不同图像尺度上的局部细粒度和全局粗粒度相互作用缺失。

3. 拟定方法

如图2所示，所提出的HTNet由变压器层和层次结构中每一级的块聚合组成。变压器层独立地对每个图像块执行自注意力机制。在低层网络中，变压器层中的自注意力函数捕捉到细粒度特征。随后，块聚合过程将小图像块聚合成更大的块，从而在同一层级的不同块之间建立交互。这种聚合使得每次块聚合后都能捕捉到粗粒度特征。需要注意的是，同一层级内的所有块共享相同的参数集。最后，我们的模型中的MLP模块被应用于最终的特征图，用于微表情分类。这种模块化和层次化的设计使得HTNet能够有效地提取并整合不同尺度和粒度级别的特征，从而提高微表情识别的性能。

3.1. 顶点帧定位和光流图提取

从起始帧和顶点帧生成的光流是描述面部区域运动位移的一种有价值的方法。这些光流在微表情识别数据集[3,28]中显示出了很有前景的结果。

为了获得光学特征图，使用起始帧索引和顶点帧索引是必不可少的。然而，在微表情数据集中，起始帧索引已经提供，因此只需从视频序列中确定顶点帧索引。为此，我们采用了D&C-RoIs方法，该方法在先前关于微表情[3]的研究中被使用，前提是数据集不提供顶点帧索引。D&C-RoIs方法有效地建立了起始帧和后续帧之间的关系，使得能够准确识别顶点帧索引，从而保证了后续微表情识别任务中光流特征的可靠提取。

$$d = \frac{\sum_{i=1}^{B} h_{1i} \times h_{h2i}}{\sqrt{\sum_{i=1}^{B} h_{1i}^2 \times \sum_{i=2}^{B} h_{2i}^2}}, \tag{1}$$

其中，B是灰度直方图中的箱数，h1是第一帧的灰度直方图，h2是当前帧。d是两帧之间LBP特征差异率。最高的是

将选择LBP特征的差异，它可以识别出发生最多面部动作的帧的索引[28,32]。

然后，我们使用起始帧和顶点帧获得光流特征图。可以如下表述光流特征图：

$$V = (u(x,y), v(x,y))|x = 1,2,\ldots..,X, y = 1,2,\ldots,Y, \quad (2)$$

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）是光流特征图V的水平和垂直分量，V = [Vx；Vy]，其中V∈RW×H×2。

在我们的方法中，我们利用光流场的一阶导数来计算光流场的变化，通常称为光学应变。光学应变提供了面部位移程度的估计，从而为微表情期间发生的细微运动提供了宝贵的见解。这种光学应变的计算使我们能够捕捉和分析复杂的面部动态，有助于准确识别微表情：

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

(3)其中)V)xx 2，$\frac{\partial V_y}{\partial y}$2）、V）yx2和$\frac{\partial V_y}{\partial x}$2是V的偏一阶导数。最后，形成三维光流特征图，并表示为Vm = [Vx；Vy；Vz]和Vm∈RW×H×3。

在我们的网络中，我们采用了一种区域特定的方法，专注于四个特定的面部区域——左眼区域、左唇区域、右眼区域和右唇区域——而不是考虑整个面部区域。为了从整个光流特征图Vm中提取这四个特定的面部光流特征图，我们使用多任务级联卷积网络（MTCNN）[33]从顶点图像中获取面部关键点坐标。然后，我们以这些面部关键点的坐标为中心，获得四个面部光流图。具体来说，左眼的光流图以左眼面部关键点为中心，同样地，右眼的光流图以右眼面部关键点为中心。随后，这四个面部光流特征图从Vm中裁剪出来，分别表示左眼光流特征图、左唇视觉特征图、右眼光流特征图和右唇特征图。每个特征图的大小为$\frac{W}{2}\times\frac{H}{2}$×3，即整个光流图像大小的一半。在提取了四个光流特征后，我们将它们结合，并将组合后的特征输入我们的HTNet进行微表情识别。整个过程如图2所示。

3.2. 变压器层

在我们的方法中，处理输入的光学流图像时，每个图像块的大小为P×P，该图像的尺寸为H×W×3。经过线性投影和分割后，每个补丁具有特征维度P×P×3。随后，这些补丁被展平，形成我们模型的输入，记作X∈Rb×Hn×n×d，其中Hn表示h中的块数。ea层级数为c，批量大小为b，序列长度为n，Hn×n = P2。每个块内应用多个变压器层，层次结构决定了所使用的变压器层数量。每个变压器层由层归一化（LN）、多头自注意力（MSA）层和前馈全连接网络（FF N）组成。为了编码空间信息，在Rd中的所有序列向量中引入了一个可训练的位置嵌入向量。这确保了空间关系和位置信息在特征表示中被有效地捕获和编码：

×W

$$YI'i=YIi?1+MSA(LN(YIi?1)) \quad YIi+1=YI'i + FFN(LN(YI'i)) \quad (4)$$

$$YI'i+1 = YIi+1 + MSA(LN(YIi+1)) \; ; YIi+2=YI'i+1 + FF N(LN(YI'i+1))$$

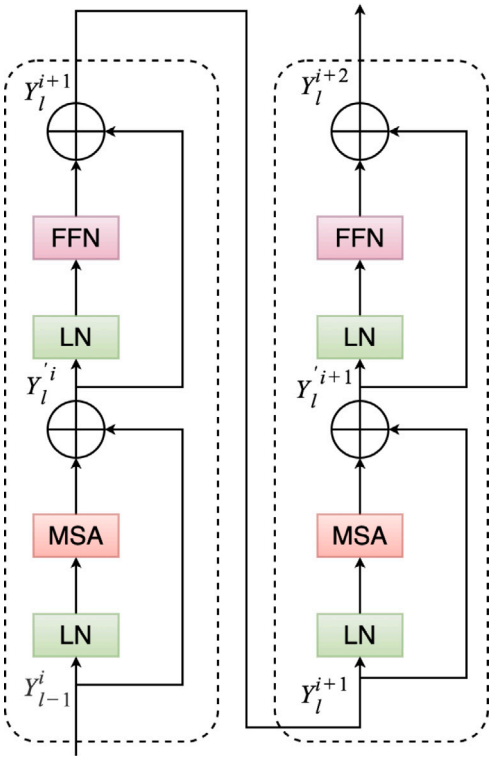


图3.每个块中将并行应用多个变压器层。层次结构决定了变压器层数量。变压器层由层归一化（LN）、多头自注意力（MSA）层和前馈全连接网络（FF N）组成。空间信息通过向所有序列向量添加可训练的位置嵌入向量来编码。

其中I = 1；2；...；L，I是每个层次i中第I个块的索引，L是每个层次中块的总数（见图3）。FF N包含两层：max（0；xW1+b）W2+b。在同一层级内的每个块i中应用多头自注意力机制。在这个自注意力组件中，输入X∈Rn×d被转换为三个部分，即查询Q、键K和值V，其中n表示序列长度，d是输入的维度。随后，在Q、K和V上使用缩放点积注意力：

$$MSA(Q;K;V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V. \quad (5)$$

LN将应用于每个区块，如下所示：

$$LN(x) = \frac{x-}{o} + , \quad (6)$$

其中，是特征的均值，是特征的标准差，o是逐元素点积，和是可学习参数。

在变压器层之后，我们采用块聚合来合并变压器层的输出。具体来说，我们通过块聚合过程将每四个小块组合成一个更大的块。

3.3. 块聚合

我们在HTNet中使用的块聚合函数与几种金字塔设计有相似之处。然而，一个显著的区别在于我们的模型对每个图像块采用局部注意力机制，而不是对整个图像采用全局注意力机制。这种方法对于提升模型性能非常有益，因为微表情识别很大程度上依赖于局部面部肌肉运动。

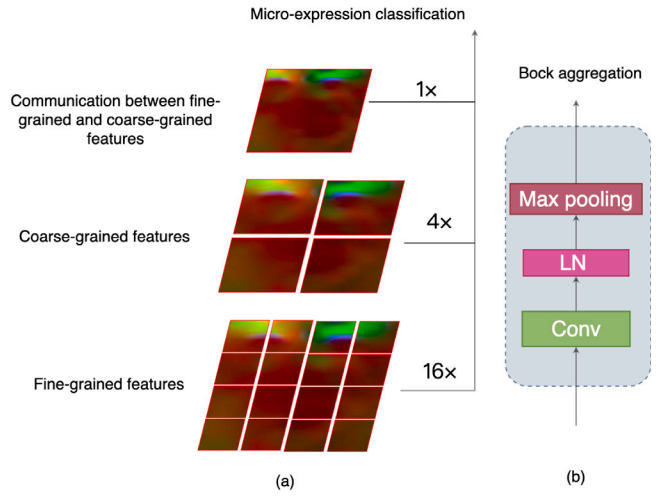


图4.块聚合包括3×3卷积层，随后是LN和3×3最大池化。在模型底部，面部光流图包含16个面部块，每个块为4×4特征图。通过在这个特征图上使用3×3卷积层，四个面部区域内的部分特征图将被合并，块大小变为2×2，对应于四个面部区域——左眼区、左唇区、右眼区和右唇区。

通过专注于特定的面部区域并利用局部注意力，我们的模型有效地捕捉了推断微表情状态的基本特征，同时忽略了无关的面部区域。

在我们的HTNet模型中，每个块独立处理光流图，并且块通信发生在块聚合期间。块聚合作用于四个相邻的块，促进它们之间的信息交换，并实现局部和全局特征的融合。具体来说，低级块聚合专注于在四个面部区域——左眼区、右眼区、左唇区和右唇区之间交换信息。这一过程提取了捕捉详细面部动态的细粒度特征。另一方面，高级块聚合促进了四个面部区域之间的全局信息交换，从而提取出捕捉更广泛面部表情的粗粒度特征。整个过程如图4所示。在层次l中，光流图像大小表示为$X^l \in R^{b \times h \times w \times d^l}$。经过块聚合后，光流图像大小变为$X^{l+1} \in R^{b \times 2h \times 2w \times d^{l+1}}$。随后，所有四个面部区域合并成一个面部特征图，记作$X \in R^{b \times H \times W \times D}$，其中$d^{l+1} > d^l$以有效保留和增强特征。

HTNet中的块聚合过程包括一个3×3卷积层，随后是层归一化（LN）和3×3最大池化。在模型的底部，面部光流图由16个面部块组成，每个块由一个4×4特征图表示。第一个3×3卷积层将四个面部区域的部分特征图合并，将块大小减少到2×2，对应左眼区域、左唇区、右眼区域和右唇区。随后，另一个3×3卷积层促进四个面部区域之间的信息交换，进一步将块大小减小到1×1，并提取完整的光学特征图。最后，提取的完整特征图被输入MLP层进行微表情分类。这一层次化过程有效地捕捉并整合了不同粒度级别的特征，有助于提高微表情识别的准确性。

3.4. 损失函数

在本文中，我们采用交叉熵损失函数来训练我们的模型。交叉熵损失L的计算可以表示为

表1
实验在SAMM [34]、SMIC [35]、CASME II [36]和CAS（ME）3 [37]数据库上进行。SAMM、SMIC和CASME II合并成一个综合数据集，这三个数据集中相同的标签被用于微表情任务。

数据库	萨姆	CASME II	山东冶金工业公司	CAS（我）3
受试者	28	24	16	100
样品	133	145	164	943
帧率	200	200	100	30
阴性	92	88	70	508
积极	26	32	51	64
意想不到的事	15	25	43	201
发作指数	✓	✓	✓	✓
偏移索引	✓	✓	✓	✓
顶点指数	✓	✓	×	✓

通过以下公式：

$$L = \sum_{i} (-w_i log(p_t^i)) \quad , p_t = p_c^y (1 - p_c)^{1-y} \tag{7}$$

其中，wi是数据集中的样本的权重，y是标签，yi∈0,1。

4. 实验

4.1. 数据集

实验在四个数据库上进行：SAMM [34]、SMIC [35]、CASME II [36]和CAS（ME）3 [37]数据库。为了确保一致性和可比性，SAMM、SMIC和CASME II被合并成一个综合数据集，在这个数据集中，这三个数据集中的相同情感标签被用于微表情识别任务。在这些数据集中，情感类别划分如下：‘积极’情感类别包括‘快乐’情感类，‘消极’情感类别包括‘悲伤’、‘厌恶’、‘轻蔑’、‘恐惧’和‘愤怒’情感类，而‘惊讶’情感类别仅包括‘惊讶’类。

SAMM [34]：SAMM数据集包含28名参与者，133个微表情和147个长视频，其中包含343个宏表情。该数据集在动作单元编码方面非常丰富，提供了全面的面部表情信息。SAMM还提供了微表情的起始、结束和顶点索引。数据集中的原始样本分辨率为2040乘以1088像素，帧率设定为每秒200帧。SAMM中图像的情感类别分为“厌恶”、“恐惧”、“轻蔑”、“愤怒”、“抑制”、“惊讶”、“快乐”和其他。分类为三个情感类别后，“负面”、“正面”和“惊讶”的数量分别为92、26和15。

CASME II [36]：CASME II数据集包含来自24名受试者的数据，共计145个样本，对应145种情绪。所有样本均使用实验室摄像头拍摄，帧率为每秒200帧。样本的原始尺寸为640×480像素。CASME II中的样本被分为“快乐”、“惊讶”、“厌恶”、“悲伤”、“恐惧”、“压抑”和其他类别。合并后，负面、正面和惊讶的情绪样本数量分别为88、32和25。CASME II中标注了情绪的起始点、结束点和顶点指数。

CAS（ME）3 [37]：CAS（ME）3是面部自发微表情数据库的第三代，以其包含深度信息和高生态效度而著称，使其成为微表情识别的宝贵资源。CAS（ME）3第A部分包括来自100名受试者的数据，共计943个样本，对应943种情绪。这些样本使用实验室相机拍摄，帧率为每秒30帧，原始分辨率为

1280 × 720像素。CAS（ME）3部分A中的样本被分类为“快乐”、“愤怒”、“恐惧”、“厌恶”、“惊讶”、“他人”和“悲伤”。负面、正面和惊讶的总数量分别为508、64和201。

SMIC[35]：SMIC-HS数据集包含来自16名受试者的数据，共计164个样本，对应164种情绪。所有样本均使用实验室相机以每秒100帧的帧率捕捉。样本的原始图像尺寸为640 × 480像素。SMIC中的样本分为“负面”、“惊讶”和“正面”。其中，“负面”、“正面”和“惊讶”的数量分别为70、51和43。SMIC中提供了起始点和结束点，但未提供顶点索引。这三个数据集的详细信息可以总结如下：

表1

4.2. 实施细节

最初，在SMIC数据集的情况下，顶点帧索引缺失，我们使用D&C-ROIs技术[38]来确定顶点帧的索引。对于SAMM、CASME II和CAS（ME）3数据集，顶点帧的真值是可用的，这简化了获取关键微表情时刻的过程。从数据集中获得起始和顶点图像后，我们使用Gunnar Farneback的[39]算法从这些图像中提取起始和顶点时间点的光流。接下来，光流图像的三个元素——水平、垂直和光学应变元素——均为28 × 28 × 3像素。随后，我们采用多任务级联卷积网络（MTCNN）[33]从顶点图像中提取面部特征坐标。这些面部特征坐标对于高精度定位特定面部区域至关重要。基于面部特征坐标，我们提取了四个重要的面部区域，即左眼、左唇、右眼和右唇的光流特征图。这四个光流特征图的大小仅为整个光流图像的一半，尺寸为14 × 14 × 3像素。通过专注于这些特定的面部区域，我们可以有效捕捉与微表情相关的面部肌肉运动。在提取了四个光学流特征图之后，我们的HTNet网络接收并组合它们。这种全面的方法确保了我们的HTNet能够准确高效地识别和分类微表情。

实验使用PyTorch和Python 3.9在Ubuntu 22.04操作系统上进行，将训练参数的学习率设置为5 × 10?5，并使用最大800个周期进行训练。

设置：微表情任务的标准评估方法是留一法（LOSO）交叉验证。LOSO交叉验证更受青睐，因为它能够公平地比较不同模型，并确保模型性能不受特定个体特征的影响。这种方法紧密模拟了现实场景，在这些场景中，人们会在各种环境和地点遇到具有不同背景和表情的个体。

性能指标：复合微表情数据集中的类别分布不平衡，不同情绪的频率不同。具体来说，惊讶、积极和消极情绪的比例大约是1：1.3：3。为了解决这个问题，我们采用未加权平均召回率（U AR)和未加权F1分数(UF 1）来报告我们的结果。

(1) 未加权F1分数(UF 1），也称为宏观平均F1分数，是一种常用于评估多类分类任务中不平衡类别分布性能的指标。为了计算UF 1，我们需要计算每个类别c在所有留一法交叉验证中的假阳性（FP）、真阳性（TP）和假阴性（FN）。

（LOSO）交叉验证。然后，可以使用公式计算每个类别的F1分数F 1c：

$$F1_c = \frac{2 \times TP_c}{2 \times TP_c + FP_c + FN_c}$$
$$UF1 = \frac{F1_c}{C} \tag{8}$$

在公式(8)中，C是类别的数量。

(2) 未加权平均召回率（U AR)：UAR是一个特别有用的评估模型有效性的指标

在存在不平衡类别比例的情况下，nc是每个类别的样本总数。U AR可以定义如下：$UAR = \frac{1}{C}\sum_c \frac{TP_c}{n_c}$ (9)

4.3. 与最新技术水平的比较

在表2和表3中，我们展示了我们的方法与之前的手工和深度学习方法在微表情数据集上的性能，即CASME II、CAS（ME）3、SMIC和SAMM。使用的评估指标是未加权F1分数（UF 1）和未加权平均召回率（U AR）。结果已报告。粗体文字用于突出显示每种数据集和度量方法中达到的最佳结果。LBP TOP [19]引入了一种利用具有六个交点的局部二值模式提取面部特征的新颖微表情识别方法。Bi-WOOF [6]从顶点图像中提取关键面部特征，他们提出了双权重光流法。该技术有效捕捉了相关的面部运动，并突出了微表情识别所需的重要信息。OFF-ApexNet [3]建议利用起始帧和顶点帧的光流场。在光流场中，获取水平和垂直特征并输入基于CNN的网络以进一步增强特征。之后，从OFF-ApexNet中提取的特征将用于微表情分类。STSTNet [28]提出使用一个三层浅层CNN模型来获得高层次的判别表示，用于分类微表情情绪。为了应对跨数据库微表情识别的挑战，Dual-Inception [40]提出了一种新颖的方法，使用两个inception网络从光流图中提取水平和垂直特征。通过将光流的水平部分输入一个inception网络，垂直部分输入另一个inception网络，他们可以独立地捕捉来自两个方向的相关模式和信息。鉴于面部情感数据集中训练样本数量有限，EMR [5]引入了两种领域适应策略。第一种策略涉及对抗训练方法。第二种策略是表达放大法。RCN [29]提出的方法建议使用较小尺寸的图像作为输入，并采用更小的架构模型，这已被证明有助于提高模型在复合数据集任务上的性能。FeatRef. [13]包括两个阶段，在第一阶段，水平inception网络和垂直inception网络将提取水平和垂直肌肉运动特征。之后，水平和垂直特征将被合并并输入到三个基于注意力的网络中，以将这些提取的特征分类为不同的微表情类别。最后，分类分支用于融合从初始模块获得的显著和区分性特征，以推断微表情。SLSTT-LSTM [31]创建了一种新的系统，用于识别快速、细微的面部表情。这是首个仅使用变压器的系统，变压器是一种不需要常规图像处理网络的模型类型。他们的系统分为三个部分：一部分学习空间和形状，另一部分观察时间上的变化，最后一部分决定面部表情是什么。

表2
LOSO协议下手工方法、深度学习方法和我们的HTNet方法在复合（全）、SMIC、CASME II和SAMM上的无权重F1分数（UF 1）和无权重平均召回率（U AR）性能。粗体表示最佳结果。

方法	全部		山东冶金工业公司		CASME II		萨姆	
	足球	乌拉圭	足球	乌拉圭	足球	乌拉圭	足球	乌拉圭
Lbp-顶部[19]	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
双动力[6]	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
AlexNet [12]	0.6933	0.7154	0.6201	0.6373	0.7994	0.8312	0.6104	0.6642
GoogLeNet [11]	0.5573	0.6049	0.5123	0.5511	0.5989	0.6414	0.5124	0.5992
VGG16 [10]	0.6425	0.6516	0.5800	0.5964	0.8166	0.8202	0.4870	0.4793
OFF-ApexNet [3]	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
STSTNet [28]	0.7353	0.7605	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810
CapsuleNet [41]	0.6520	0.6506	0.5820	0.5877	0.7068	0.7018	0.6209	0.5989
双发夹层[40]	0.7322	0.7278	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663
EMR [5]	0.7885	0.7824	0.7461	0.7530	0.8293	0.8209	0.7754	0.7152
RCN [29]	0.7432	0.7190	0.6326	0.6441	0.8512	0.8123	0.7601	0.6715
FeatRef [13]	0.7838	0.7832	0.7011	0.7083	0.8915	0.8873	0.7372	0.7155
SLSTT-LSTM [31]	0.816	0.790	0.740	0.720	0.901	0.885	0.715	0.643
HTNet（我们的）	**0.8603**	**0.8475**	**0.8049**	**0.7905**	**0.9532**	**0.9516**	**0.8131**	**0.8124**

4.3.1. 与手工方法相比

表2展示了不同方法的比较分析，包括使用基于外观和几何技术提取面部特征的LBP-TOP和Bi-WOOF。这两种方法均采用SVM作为分类器。相比之下，我们提出的方法HTNet在复合数据集上的UF1和UAR有显著提升。具体而言，UF1从0.6296提高到0.8603，UAR从0.6227提高到0.8475，提升了超过20%。此外，HTNet在CASME II、SMIC和SAMM数据集上始终优于手工设计的方法（LBP-TOP和Bi-WOOF）。这些结果强调了HTNet在解决领域转换和在微表情识别任务中实现优于手工和深度学习方法的性能方面的有效性。

4.3.2. 与深度学习方法相比

在表2和表3中，我们的HTNet显著优于大多数深度学习方法。如表2所示，HTNet在完整复合数据集上分别达到了0.8603和0.8475的UF1和UAR，相比之前最先进的方法提高了大约5%。分析表2，我们观察到ALexNet、GoogLeNet和VGG16在完整复合数据集上的UF1和UAR分别为0.6933和0.7154、0.5573和0.6049、以及0.6425和0.6516。这些更深的深度学习方法在完整复合数据集上的表现不如其他较浅的方法（STSTNet、双子网络、RCN）。这一结果的原因在于这三种方法使用了更深的网络，可能会因为训练样本数量有限而引入噪声信息。尽管较浅的网络相比VGG16取得了优越性能，但一个关键的研究问题是如何从微表情中提取显著且具有区分性的特征。FeatRef能够从微表情中提取显著且具有区分性的特征，因此其方法实现了UF1和UAR分别为0.7838和0.7832。然而，FeatRef忽略了微表情识别中的局部和全局时空模式。目前仍能与我们方法竞争的一种方法是SLSTTLSTM [31]。他们使用变换器在不同时间序列中提取图像特征，并使用LSTM沿不同时间合并这些特征。尽管SLSTT-LSTM对处理序列中不同图像之间的变化敏感，但它可能忽略了单个光流图像中的空间信息，以及光流图像中不同区域之间的关系缺失。相比之下，HTNet专注于四个面部区域而非整个面部区域。通过在每个面部区域引入局部自注意力机制，变换层可以集中于定位细微的肌肉收缩。此外，聚合层有助于学习不同分辨率的光流之间的交互

表3
LOSO协议下CAS（ME）3 Part A中深度学习方法和我们的HTNet方法的无权重F1分数（UF1）和无权重平均召回率（UAR）性能。粗体表示最佳结果。

方法	CASME3A部分	
	足球	乌拉圭
AlexNet [12]	0.257	0.2634
STSTNet [28]	0.3795	0.3792
RCN [29]	0.3928	0.3893
FeatRef [13]	0.3493	0.3413
HTNet (Ours)	**0.5767**	**0.5415**

特征映射。因此，HTNet在完整的复合数据集上比以前的方法取得了更好的性能。

在表3中，我们对CAS（ME）3 Part A进行了实验，并报告了深度学习方法的未加权F1分数（UF1）和未加权平均召回率（UAR），包括我们的HTNet，在留一受试者（LOSO）协议下。HTNet显著优于先前的方法，证明了其在微表情识别中的有效性。

4.4. 消融研究

本节对HTNet模型中各种参数的影响进行了深入分析，研究了块大小、隐藏维度、变压器中的头数以及每个层次的变压器层数的影响。初始实验设置包括底层块大小为7×7，隐藏维度为256，变压器中有三个头，每个层次的变压器层数为（2；2；8）。对于每次消融实验，我们修改一个特定参数，同时保持其他设置不变。每种方法的评估使用UAR指标和UF1指标进行。

4.4.1. 块大小的影响

我们进行了实验，研究不同块大小对复合微表情数据集整体准确性的影响，包括SMIC、SAMM和CASME II。块大小指的是HTNet模型中考虑的面部区域的大小。我们将底层块大小从5调整到10，中间层块大小是底层大小的两倍，顶层块大小是底层大小的四倍。表4中的结果显示，块大小的选择显著影响了模型的性能。较小的块大小可能导致表现不佳。

表4
研究不同块大小对复合数据集-SMIC、SAMM和CASME II的影响。报告了复合数据集的未加权F1分数（UF1）和未加权平均召回率（UAR）性能。

# Block 尺寸（底部水平）		5 × 5	6 × 6	7 × 7	8 × 8	9 × 9	10 × 10
# Block 尺寸（中层）		10 × 10	12 × 12	14 × 14	16 × 16	18 × 18	20 × 20
# Block 大小（顶级）		20 × 20	24 × 24	28 × 28	32 × 32	36 × 36	40 × 40
全部	足球	0.837	0.8271	**0.8603**	0.85	0.8511	0.8546
全部	乌拉圭	0.8213	0.8037	0.8475	**0.846**	0.8445	0.8383
山东冶金工业公司	足球	0.7556	0.7394	**0.8049**	0.7833	0.7753	0.8008
山东冶金工业公司	乌拉圭	0.7469	0.7214	**0.7905**	0.7792	0.7708	0.7892
萨姆	足球	**0.8237**	0.7903	0.8131	0.7909	0.8168	0.8068
萨姆	乌拉圭	0.8033	0.775	**0.8124**	0.79	0.8088	0.7812
卡斯梅 微光	足球	0.9422	0.9482	0.9532	**0.964**	0.9722	0.957
卡斯梅 微光	乌拉圭	0.9308	0.9317	0.9516	0.962	**0.9658**	0.945

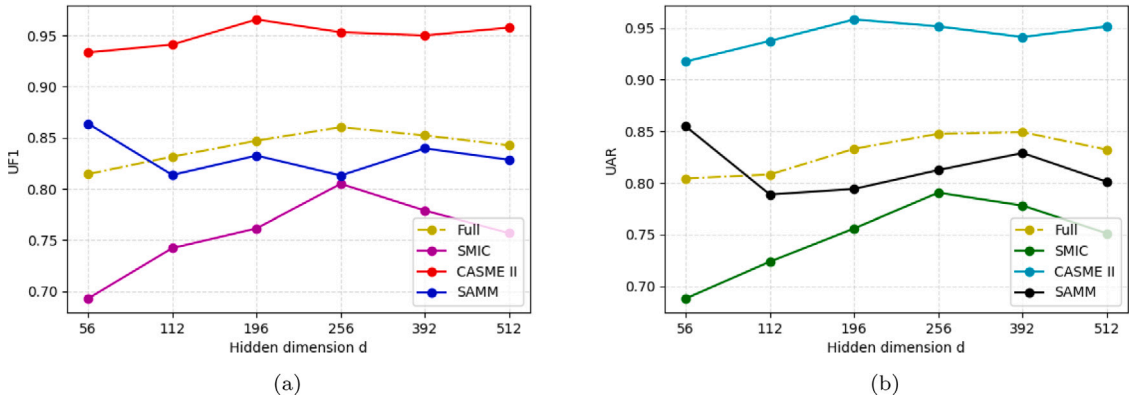


图5.我们研究维度数量如何影响复合数据集的准确性——SMIC、SAMM和CASME II。复合数据集的未加权F1分数（UF1）并报告了未加权平均召回率（UAR）性能。

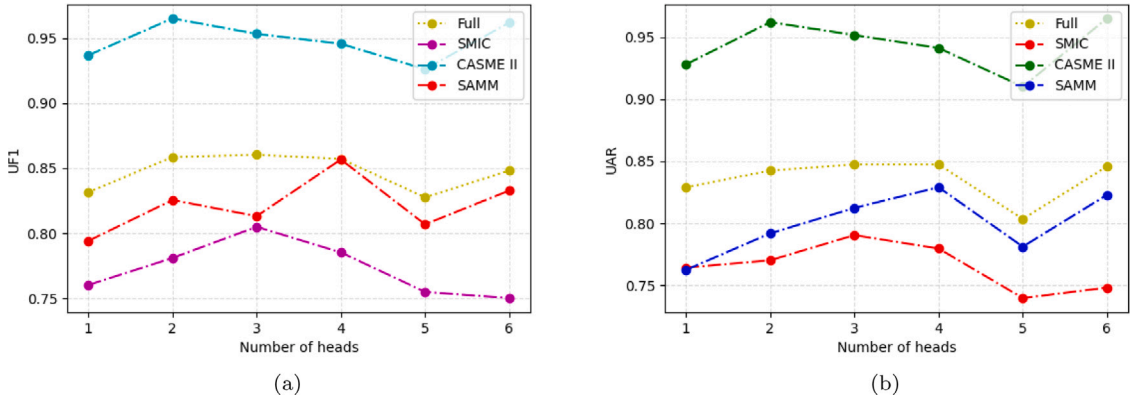


图6.我们研究了变压器层的头数对复合数据集-SMIC、SAMM和CASME II的准确性的影响。复合数据集的未加权F1分数报告了（UF1）和未加权平均召回率（UAR）性能。

他们可能会遗漏一些关键的面部部位，而较大的块大小通常表现更好。然而，如果块大小变得太大，一些面部区域可能会重叠，例如左眼区域与右眼区域重叠，这可能对模型的表现产生负面影响。因此，在块大小之间找到一个最佳的平衡是确保在微表达识别任务中获得最佳性能的关键。

4.4.2. 尺寸的影响

我们研究了维度数量如何影响复合数据集的准确性——SMIC、SAMM和CASME II。图5报告了复合数据集的未加权F1分数（UF1）和未加权平均召回率（UAR）。较小的隐藏维度表现较差，因为小的隐藏维度难以编码光流特征图。然而，使用较大的隐藏维度会导致过拟合问题。隐藏维度256的完整UF1得分表现最好，约为0.86。

4.4.3. 头数的影响

我们研究了变压器层的头数对复合数据集——SMIC、SAMM和CASME II的准确性的影响。图6报告了复合数据集的未加权F1分数（UF1）和未加权平均召回率（UAR）。图6(a)显示，变压器层中的三个头将表现出最佳性能。

4.4.4. 不同变压器块数量的影响

表5展示了使用多个数据集评估变压器层数对我们的HTNet模型准确性的影响。为了探索层数的影响，

表5
研究变压器层数对复合数据集的影响。报告复合数据集的未加权F1分数（UF1）和未加权平均召回率（UAR）性能。

#变压器层数		(2,2,2)	(2, 2, 4)	(2, 2, 6)	(2, 2, 8)	(2, 2, 10)	(2, 2, 12)
全部	足球	0.8105	0.8297	0.8316	**0.8603**	0.8464	0.8499
全部	乌拉圭	0.7848	0.8029	0.8183	**0.8475**	0.8376	0.8207
山东冶金工业公司	足球	0.748	0.7247	0.744	**0.8049**	0.76	0.7561
山东冶金工业公司	乌拉圭	0.7363	0.7065	0.734	**0.7905**	0.756	0.7345
萨姆	足球	0.7984	0.8137	0.7732	0.8131	0.8571	**0.8682**
萨姆	乌拉圭	0.75	0.7849	0.7742	0.8124	**0.8453**	0.8419
CASME II	足球	0.8845	0.9647	**0.9722**	0.9532	0.9492	0.95
CASME II	乌拉圭	0.8588	0.9554	**0.9658**	0.9516	0.9346	0.9412
#参数（百万）		149.57	245.88	342.20	438.51	534.82	631.13
#训练时间（秒）		3767	4836	5906	6942	8085	9147

表6
报告了未加权F1分数（UF 1）和未加权平均召回率（U AR），以分析HTNet的层次数量。

#级别数	全部		山东冶金工业公司		CASME II		萨姆	
	足球	乌拉圭	足球	乌拉圭	足球	乌拉圭	足球	乌拉圭
1	0.7089	0.6747	0.6327	0.6220	0.8567	0.8398	0.5948	0.5490
2	0.8588	0.8375	0.7923	0.7755	0.9607	0.9620	0.8267	0.7938
3	0.8500	0.8510	0.7820	0.7821	0.9526	0.9430	0.8137	0.8251

在顶层中，我们改变变压器层的数量，从2到12。观察到，变压器层数的增加会导致模型参数和训练时间相应增长。较少的变压器层数可能无法有效编码光流特征。相反，更多的层数可能会引入过拟合问题，影响模型性能。我们的实验表明，在底层和中间层使用两个变压器层，以及在顶层使用适量的层，可以得到我们HTNet模型的最佳性能。

4.4.5. 不同层次结构数量的影响

消融研究，如表6所示，将特别探讨不同层次结构对HTNet性能的影响。报告了复合数据集上的未加权F1分数（UF1）和未加权平均召回率（UAR），以展示我们模型中层次方法的有效性。这为不同层次数量对微表情识别准确性的影响提供了清晰的见解。在我们的HTNet中没有层次结构时，模型在复合数据集上的UF1得分为0.7089，UAR得分为0.6747；在SMIC数据集上的UF1得分为0.6327，UAR得分为0.6222；在CASME II数据集上的UF1得分为0.8567，UAR得分为0.8398；在SAMM数据集上的UF1得分为0.5948，UAR得分为0.5490。但增加一个层次结构后，模型可以将复合数据集的UF1得分和UAR得分从0.7089到0.8588和0.6747到0.8375，SMIC数据集的UF1分数和UAR分数从0.6327到0.7923和0.6220到0.7755，CASME II数据集的UF1分数和UAR分数从0.8567到0.9607和0.8398到0.9620，SAMM数据集的UF1分数和UAR分数从0.5948到0.8267和0.5490到0.7938。这表明层次结构将提高识别微表情的能力，使其更加准确。

5. 我们所提出方法的定性和定量分析

图7(a)展示了HTNet的混淆矩阵，呈现了在完整复合数据集中每个情感类别的准确率。值得注意的是，在这些数据集（SMIC、SAMM和CASME II）中，HTNet分别达到了0.94、0.81和0.80的负、正和惊讶类别的准确率。负类别较高的准确率可以归因于这三个数据集中该类别的训练样本数量较多。然而，挑战也随之而来。

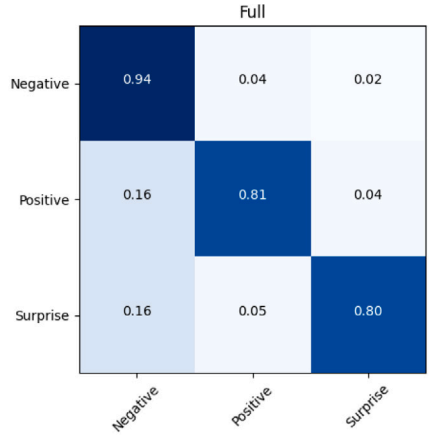
在区分SMIC和SAMM数据集中的惊喜和负面情绪类别。这些数据集中有限的训练样本可能导致这两个类别之间偶尔出现错误分类。因此，CASME II数据集中的负面、正面和惊讶类别达到了相对较高的准确率，每个都超过了90%。错误分类极少，进一步验证了HTNet在该数据集上的有效性。此外，SMIC数据集使用较低的帧率捕捉图像，这引入了背景噪声，如闪烁的灯光、阴影和光照变化。这些因素可能会影响SMIC数据集中正面和惊讶情绪分类的准确性，分别达到了76%和72%。

图7(e)给出了CAS（ME）3的混淆矩阵。
从混淆矩阵中，我们观察到负面情绪类别在三个情绪类别中达到了最高的准确率。这一结果与复合数据集中的表现一致，在这些数据集中，负面类别同样表现出最高的准确率。另一方面，正面情绪类别的准确率最低，约为0.15。大多数阳性样本被错误分类为阴性。相反，在复合数据集中，只有少量样本被错误分类为阴性。

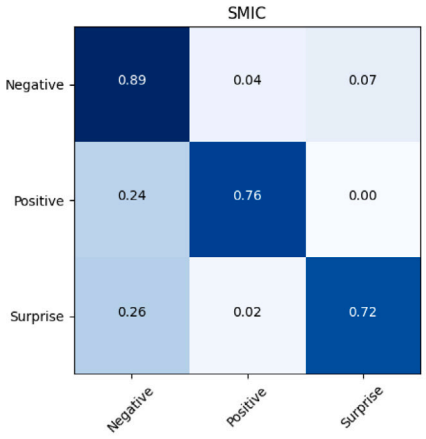
6. 结论

在本文中，我们提出了一种分层变换器架构，用于提取微表情识别中重要的四个面部区域特征。与以往的深度学习方法相比，该模型的关键优势之一在于其分层架构能够提取多尺度的特征，从而提高表情识别的准确性。此外，HTNet模型基于transformer架构，该架构在自然语言处理任务中表现出色。这种架构非常适合微表情识别，因为它不仅能够建模面部表情不同部分之间的复杂依赖关系，同时还在计算上非常高效。

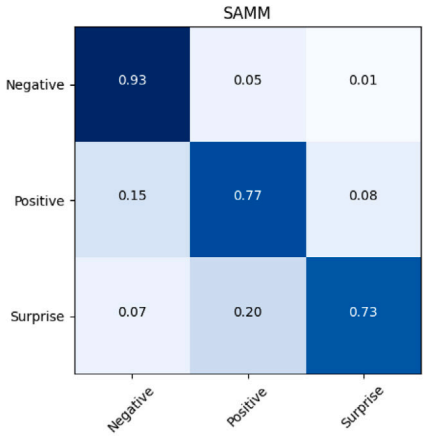
尽管有这些优点，HTNet模型需要大量的训练数据，在某些情况下可能难以获得。该模型也可能对光线、姿势和其他环境因素的变化敏感，这些因素可能会影响微表情的外观。



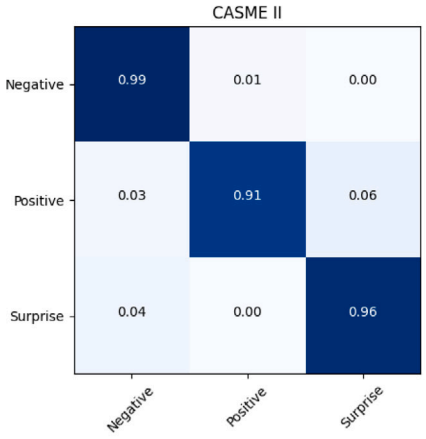
(a) Full dataset



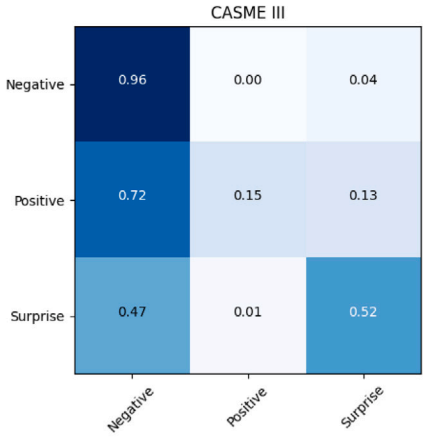
(b) SMIC dataset



(c) SAMM dataset



(d) CASME II dataset



(e) CAS(ME)^3 dataset

图7.在复合数据库（SMIC、SAMM、CASME II）和CAS（ME）3数据集上使用3个类别提出的HTNet的混淆矩阵。

为了克服这些限制，未来的工作可以探索提高HTNet模型效率和鲁棒性的方法。可以使用数据增强技术来增加训练数据量，提高模型在不同环境条件下的泛化能力。还可以利用迁移学习，借鉴相关任务上的预训练模型，从而提升HTNet模型的效率。通过进一步的发展和优化，HTNet该模型可能被证明是准确识别各种应用中微表情的有价值的工具，包括谎言检测、情绪识别和心理健康评估。

CRediT作者贡献声明

王志峰：撰写-审阅与编辑，撰写-初稿，可视化，验证，资源，方法学，

调查、正式分析、概念化。张凯浩：撰写-审阅与编辑、监督、资源、方法论、调查、概念化。罗文涵：撰写-审阅与编辑、监督、方法论、调查、概念化。拉梅什·桑卡纳拉亚纳：撰写-审阅与编辑、撰写-原始草稿，监督，方法学，正式分析，概念化。

利益冲突声明

作者声明他们没有已知的竞争性经济利益或个人关系，这些关系可能会对本文报告的工作产生影响。

数据可用性

应要求提供数据。

致谢

我谨向我的导师表示感谢，感谢他们给予的宝贵讨论、帮助和支持。这项工作得到了澳大利亚政府研究培训计划奖学金的部分资助。

参考文献

[1] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, W. Zhang, FERV39k：用于视频中面部表情识别的大规模多场景数据集，见：IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第20922-20931页。
[2] S. Thuseethan, S. Rajasegarar, J. Yearwood, EmoSeC：从场景上下文识别情绪，Neurocomputing 492 (2022) 174-187。
[3] Y.S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, L.-K. Tan, OFF-ApexNet在微表情识别系统上的应用，Signal Process., Image Commun.74 (2019) 129-139。
[4] Y. Gan, J. See, H.-Q. Khor, K.-H. Liu, S.-T. Liong, Needle in a haystack： Spotting and recognizing micro-expressions in the wild, Neurocomputing 503 (2022) 283-298。
[5] Y. Liu, H. Du, L. Zheng, T. Gedeon, 一种神经微表情识别器，见：2019年第14届IEEE国际自动面部和手势识别会议，FG 2019，IEEE，2019年，第1-4页。
[6] S.-T. Liong, J. See, K. Wong, R.C.-W. Phan, 少即是多：利用顶点帧从视频中识别微表情，信号处理，图像通信。62 (2018) 82-92。
[7] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, X. Fu, 自发微表情识别的主要方向平均光流特征，IEEE Trans.Affect.Comput.7(4) (2015) 299-310。
[8] S. Happy, A. Routray, 用于微表情识别的模糊光流方向直方图，IEEE Trans. Affect.Comput.10(3) (2017) 394-406。
[9] S.-T. Liong, R.C.-W. Phan, J. See, Y.-H. Oh, K. Wong, 基于光学应变的微妙情绪识别，见：2014年国际智能信号处理与通信系统研讨会，ISPACS，IEEE，2014，pp.180－184。
[10] A. Sengupta, Y. Ye, R. Wang, C. Liu, K. Roy, 深入研究脉冲神经网络：VGG和残差架构，Front. Neurosci. 13 (2019) 95。
[11] P. Ballester, R.M. Araujo, 关于应用于草图的GoogLeNet和AlexNet的性能，第30届AAAI人工智能会议，2016年。
[12] H. Zhang, H. Zhang, 基于深度学习的微表情识别综述，2022年国际神经网络联合会议，IJCNN，IEEE，2022，第01-08页。
[13] L. Zhou, Q. Mao, X. Huang, F. Zhang, Z. Zhang, 特征细化：一种用于微表情识别的特定表达特征学习和融合方法，Pattern Recognit. 122 (2022) 108275。
[14] B. Pan, K. Hirota, Z. Jia, Y. Dai, 多模态情感识别数据集、预处理、特征和融合方法综述，Neurocomputing (2023) 126866。
[15] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, P.Xiao, 基于多任务和集成学习的多特征视觉音频情感识别，神经计算391 (2020) 42-51。
[16] A. Bhandari, N.R. Pal, 边缘是否有助于卷积神经网络在情绪识别中的应用？神经计算433 (2021) 162-168。
[17] G. Zhao, M. Pietikainen, 利用局部二进制模式进行动态纹理识别及其在面部表情中的应用，IEEE Trans. Pattern Anal.Mach.Intell.29(6) (2007) 915-928。
[18]王思杰、严伟军、李翔、赵刚、付翔，基于张量独立颜色空间动态纹理的微表情识别，2014年第22届国际模式识别会议论文集，IEEE，2014年，第4678-4683页。
[19] Y. Wang, J. See, R.C.-W. Phan, Y.-H. Oh, LBP with six intersection points：减少LBP-top中微表情识别的冗余信息，见：亚洲计算机视觉会议，Springer，2014，pp.525-537。
[20]王思杰、严伟军、李翔、赵刚、付翔，基于张量独立颜色空间动态纹理的微表情识别，第22届国际模式识别会议论文集，IEEE，2014年，第4678-4683页。
[21] X.李、J.余、S.Zhan，基于深度学习的自发面部微表情检测，见：2016年IEEE第13届国际信号处理会议，ICSP，IEEE，2016年，第1130-1134页。
[22]卢志，罗志，郑浩，陈杰，李伟，基于德拉诺伊的微表情识别时间编码模型，见：计算机视觉-ACCV 2014工作坊：新加坡，新加坡，2014年11月（2014）1-2，修订精选论文，第二部分12，施普林格出版社，2015年，第698-711页。
[23] L. Lei, T. Chen, S. Li, J. Li，基于面部图表示学习和面部动作单元融合的微表情识别，IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第1571-1580页。
[24] K. Zhang, Y. Huang, H. Wu, L. Wang，基于深度学习特征的面部微笑检测，第3届IAPR亚洲模式识别会议论文集，ACPR，IEEE，2015，pp.534-538。
[25] K. Zhang, Y. Huang, Y. Du, L. Wang，基于深度进化时空网络的面部表情识别，IEEE Trans. Image Process.26(9) (2017) 4193-4203。
[26] W. Niu, K. Zhang, D. Li, W. Luo，通过潜在表达放大实现弱表达识别的四人组GAN，Knowl。基于系统的251 (2022) 109304。
[27] Z. Xia, X. Hong, X. Gao, X. Feng, G. Zhao，用于识别自发微表情的时空循环卷积网络，IEEE Trans.Multimed.22(3) (2019) 626-640。
[28] S.-T. Liong, Y.S. Gan, J. See, H.-Q. Khor, Y.-C. Huang，浅层三流三维CNN（STSTNet）用于微表情识别，见：2019年第14届IEEE国际自动面部和手势识别会议，FG 2019，IEEE，2019年，第1-5页。
[29] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, G. Zhao，基于模型和数据缩减的复合数据库微表情识别，IEEE Trans. Image Process.29 (2020) 8590－8605。
[30] A.J.R. Kumar, B. Bhanu，基于图注意力卷积网络的地标关系的微表情分类，in：IEEE/CVF计算机视觉和模式识别会议论文集，2021年，pp.1511-1520。
[31]L.Zhang, X.Hong, O.Arandjelovic, G.Zhao，基于短程和长程关系的时空变换器用于微表情识别，IEEE Trans.Affect.Comput.13(4) (2022) 1973-1985。
[32] S.-T. Liong, J. See, K. Wong, A.C. Le Ngo, Y.-H.哦, R. Phan，微表情数据库中自动顶点框架识别，见：2015年第三届IAPR亚洲模式识别会议，ACPR，IEEE，2015年，第665-669页。
[33] E. Jose, M. Greeshma, M.T. Haridas, M. Supriya，基于facenet和mtcnn的jetson tx2人脸识别监控系统，2019年第5届国际高级计算与通信系统会议，ICACCS，IEEE，2019，pp.608－613。
[34] A.K. Davison, C. Lansley, N. Costen, K. Tan, M.H. Yap, Samm：自发的微面部运动数据集，IEEE Trans.Affect.Comput.9(1) (2016) 116-129。
[35]X.Li, T.Pfister, X.Huang, G.Zhao, M.Pietikainen，自发的微表情数据库：诱导、收集和基线，in： 2013年第十届IEEE国际自动面部和手势识别会议和研讨会，FG，IEEE，2013，pp.1-6。
[36] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, CASME II：改进的自发微表情数据库和基线评估，PLoS One 9 (1) (2014) 86041。
[37]李杰，董志，陆山，王世军，严伟军，马勇，刘勇，黄超，徐。Fu，CAS（ME）3：具有深度信息和高生态有效性的第三代面部自发微表情数据库，IEEE模式分析与机器智能学报（01）（2022）1。
[38] S.-T. Liong, J. See, K. Wong, A.C. Le Ngo, Y.-H. Oh, R. Phan，微表情数据库中自动顶点框检测，第3届IAPR亚洲模式识别会议论文集，ACPR，IEEE，2015，pp.665－669.
[39] C.-W. Chang, Z.-Q. Zhong, J.-J. Liou, A FPGA实现Farneback光流通过高层次合成，FPGA‘19，计算机协会，纽约，NY，美国，2019年，第309页。

[40] L. Zhou，Q. Mao，L. Xue，跨数据库微的双概念网络
表达识别，第2019年第14届IEEE国际会议
自动面部和手势识别，FG 2019，IEEE，2019，第1-5页。
[41] N. Van Quang，J. Chun，T. Tokuyama，用于微表情识别的胶囊网络
，收录于：2019年第14届IEEE国际自动面部与手势会议
认可，FG 2019，IEEE，2019，第1-7页。

王志峰于2021年在澳大利亚国立大学获得计算机视觉与机器学习硕士学位。目前，他正在澳大利亚首都领地堪培拉的澳大利亚国立大学工程与计算机科学学院攻读博士学位。他的研究重点是计算机视觉和机器学习，特别是用于面部和情感识别的深度学习。

张凯浩（IEEE研究生会员）目前在澳大利亚国立大学工程与计算机科学学院攻读博士学位。他在国际会议和期刊上发表了超过20篇引用论文，包括CVPR、ICCV、ECCV、NeurIPS、AAAI、ACMMM、IJCV、IEEE图像处理汇刊（TIP）和IEEE多媒体汇刊（TMM）。他的研究兴趣集中在计算机视觉和深度学习。

罗文涵是中山大学的副教授兼博士生导师，研究方向包括可信人工智能和创意人工智能。在成为大学教职人员之前，他曾任腾讯应用研究科学家，利用计算机视觉和机器学习技术解决实际问题。在此之前，他在加利福尼亚州帕洛阿尔托的亚马逊（A9）工作，开发了深度模型以提升视觉搜索体验。更早之前，他在腾讯人工智能实验室担任研究科学家。他于2016年在英国伦敦帝国理工学院获得博士学位，2012年在中国科学院自动化研究所获得工学硕士学位，2009年在中国华中科技大学获得理学士学位。他发表了超过60篇同行评审的论文，其中40多篇发表在顶级会议和期刊上，如ICML、CVPR、ICCV、ECCV、AAAI、ACL、ACMMM、ICLR、TPAMI、AI、IJCV、TIP，其中有2篇是ESI高被引论文。他曾获得2019年CVPR最佳论文提名，并荣获2022年ACM中国新星奖（广州分会）。

拉梅什·桑卡兰纳亚纳获得了印度科学研究院的理学硕士学位和加拿大阿尔伯塔大学的哲学博士学位。他目前担任澳大利亚国立大学计算机科学研究所副所长（教育合作），位于澳大利亚堪培拉。他的研究兴趣包括信息检索、以人为中心的计算和软件工程。他是ACM的成员。