E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)07-2420-17

论文引用格式:Shen T S, Chi J, Wang Y B, Lei Y L and Xu M. 2025. Multi-branch micro-expression recognition method based on multi-type features and multi-attention mechanism. Journal of Image and Graphics, 30(7):2420-2436(沈天舒, 迟静, 王雁冰, 雷雁蕾, 徐铭. 2025. 融合多类型特征和多注意力机制的多分支微表情识别. 中国图象图形学报, 30(7):2420-2436)[DOI:10.11834/jig.240440]

融合多类型特征和多注意力机制的多分支微表情识别

沈天舒,迟静*,王雁冰,雷雁蕾,徐铭

1. 山东财经大学计算机与人工智能学院,济南　250014; 2. 山东省数字经济轻量智算与可视化重点实验室,济南　250014

摘　要:**目的**　针对微表情样本不足、难以学习到多类型特征、识别精度低的问题,提出一个由深度卷积神经网络与浅层3DCNN(3D convolutional neural network)组成的、融入多个注意力模块的三分支网络。**方法**　该网络在深度卷积神经网络中引入深度增强通道注意力模块ECANet(efficient channel attention network),在浅层3DCNN中提出新的面部特征增强注意力模块,通过3个分支分别提取和处理人脸特征、光流特征和光应变特征并逐层融合,从而能够充分利用各层特征中隐含的不同信息,减少识别过程中人脸细微特征的丢失,显著提高了微表情识别的精确度。此外,本文首次将IM Loss(information maximizing loss)函数引入微表情识别模型,并与Focal Loss函数结合,使模型能够更有效地处理难以分类的样本,有效提高了模型对于难样本的识别能力。**结果**　在CASME Ⅱ(Chinese Academy of Sciences micro-expression database Ⅱ)、SMIC(spontaneous micro-expression corpus)、SAMM(spontaneous actions and micro-movements)微表情数据集和MEGC2019(micro-expression grand challenge 2019 composite database)复合微表情数据集上的实验表明,所提方法在微表情识别方面优于传统方法和现有的深度学习方法。**结论**　本文提出的三分支网络显著提高了微表情识别的精确度,有效解决了微表情样本不足、难以学习到多类型特征的问题。

关键词:微表情识别;特征融合;注意力机制;光流;深度学习

Multi-branch micro-expression recognition method based on multi-type features and multi-attention mechanism

Shen Tianshu, Chi Jing*, Wang Yanbing, Lei Yanlei, Xu Ming

1. School of Computer and Artificial Intelligence, Shandong University of Finance and Economics, Ji'nan　250014, China;

2. Shandong Key Laboratory of Lightweight Intelligent Computing and Visualization for Digital Economy, Ji'nan　250014, China

Abstract: **Objective**　Micro-expressions are involuntary, brief facial movements that occur when individuals experience certain emotions, often revealing their genuine feelings despite effort to conceal them. Recognizing these expressions is valuable in a variety of real-world applications. Traditional methods for micro-expression recognition typically rely on hand-

收稿日期:2024-08-15;**修回日期**:2024-12-07;**预印本日期**:2024-12-14

*通信作者:迟静　chijing@sdufe.edu.cn

基金项目:国家自然科学基金项目(61772309);山东省中央引导地方科技发展资金项目(YDZX2023079);济南市"新高校20条"科研带头人工作室项目(2021GXRC092);山东省高等学校青创科技支持计划项目(2020KJN007);山东省重点研发计划资助(2024TSGC0118)

Supported by: National Natural Science Foundation of China (61772309); Shandong Province Central Government-Guided Local Science and Technology Development Fund Project (YDZX2023079); Jinan City "New Universities 20 Articles" Scientific Research Leaders Studio (2021GXRC092); Shandong Province Higher Education Institutions Youth Innovation Science and Technology Support Plan (2020KJN007); Shandong Provincial Key R&D Program, China (2024TSGC0118)

crafted features combined with classical machine learning techniques, which extract feature descriptors from original images. However, these methods frequently struggle with subtle and fast spatiotemporal changes inherent in micro-expressions, leading to suboptimal recognition accuracy. Optical flow is a common technique used for micro-expression recognition because it captures small motion changes and focuses on spatiotemporal information. Despite their advantages, optical flow-based methods are computationally expensive, requiring substantial resources for video processing. In addition, issues such as unclear boundaries between expression categories and limited sample sizes in datasets challenge the performance of current deep learning models, such as the 3D convolutional neural network (3DCNN), which may struggle with feature learning, generalization, and model fitting. To address these issues, this study proposes a three-branch network that is designed to overcome the challenges posed by insufficient micro-expression data, the difficulty in learning multilevel features, and low recognition accuracy. The proposed network integrates a deep convolutional neural network (CNN) and shallow 3DCNN with multiple attention mechanisms. **Method** The network incorporates an enhanced channel attention module (ECANet) within deep CNN, allowing for a more effective selection of key feature channels while reducing interference from irrelevant data. In addition, a facial feature enhancement attention module is proposed for shallow 3DCNN, enhancing the model's ability to detect subtle facial expression changes. The network uses three distinct branches to extract and process facial, optical flow, and optical strain features, which are progressively fused across layers. This approach maximizes the information captured at each layer, minimizing the loss of fine details and significantly improving recognition accuracy. Moreover, this study is the first to introduce the information maximizing loss function into micro-expression recognition in combination with the focal loss function. This weighted combination helps the model to better handle challenging samples, improving recognition accuracy for difficult cases. The proposed method consists of three major stages: preprocessing, feature extraction and fusion, and micro-expression classification. During the preprocessing stage, the apex frame of each video sequence is located to identify the peak of the expression. Optical flow and strain maps are extracted from the apex and onset frames. During the feature extraction and fusion stage, the first branch network utilizes ResNet-18 with ECANet to extract facial features, while the second and third branch networks, which are shallow 3DCNN models with the face-enhanced attention module, extract features from optical flow and strain maps. The features from the three branches are then fused layer-by-layer to enhance feature representation. During the final classification stage, the softmax function is applied to predict the probabilities of each feature that belongs to one of three micro-expression categories: negative, positive, or surprise. The class with the highest probability is chosen as the predicted result. **Result** To minimize biases during model training, this study uses leave-one-subject-out cross-validation on the Chinese Academy of Sciences Micro-expression (CASME) Ⅱ, SMIC, SAMM, and 2019 Micro-expression Grand Challenge datasets. In this approach, data from one subject are reserved for testing, while the remaining data are used for training, ensuring that no information from the test subject is included during training. The evaluation metrics used include unweighted average recall (UAR) and unweighted F1-score (UF1). The experimental results show that the proposed method achieves the highest UF1 and UAR scores on the CASME Ⅱ dataset, with scores of 0.984 3 and 0.985 7, respectively. On the SMIC dataset, the UF1 and UAR scores are 0.767 1 and 0.752 5, respectively, which also outperform existing methods. Ablation studies further confirm the effectiveness and robustness of the proposed approach. **Conclusion** The proposed micro-expression recognition model effectively reduces the loss of subtle facial features, enhances recognition accuracy, and captures key facial features while minimizing redundant information. By introducing specialized loss functions, the model also improves its ability to accurately recognize difficult samples. Compared with other mainstream models, the proposed approach achieves state-of-the-art recognition performance, and the results of the ablation studies validate the robustness of the method. In conclusion, the proposed method significantly improves the efficiency and accuracy of micro-expression recognition.

Key words: micro-expression recognition; feature fusion; attention mechanism; optical flow; deep learning

0 引 言

面部表情是一个人表达情绪状态的一种方式，而人往往会隐藏自己真实的情绪。心理学认为真实的情绪有时会以一种非常短暂和快速的面部表情表达出来，通常会持续1/25～1/3 s（Yan等，2013），并且出现在面部的特定区域，这种表情称为微表情（micro-expression，ME）（Ekman等，2003）。与宏观表情相比，微表情是一种人们无意间触发的面部情绪，自发且不可控制，可以反映人们的真实感受，因此对微表情进行研究具有重要的作用，在公安、司法刑侦和日常生活等领域有着广泛的应用。

微表情识别有3个显著特点：1）持续时间短，通常持续1/25～1/3 s；2）动作幅度小，变化强度低；3）局部运动只出现在固定的面部区域。目前的微表情识别方法一般是对数据预处理后进行特征提取，然后基于提取的特征对微表情分类。

在特征提取方面，传统方法大多通过特征描述子对原始图像进行特征提取（Huang等，2016），例如，Zhao和Pietikainen（2007）提出的三正交平面局部二值模式（local binary pattern from three orthogonal planes，LBP-TOP）通过在3个正交平面上计算LBP特征形成最终特征向量。基于此，Huang等人（2015，2016）提出时空局部二值模式与积分投影（spatio-temporal local binary pattern with integral projection，STLBP-IP）和完全局部量化模式（spatio-temporal completed local quantization pattern，STCLQP）。Zong等人（2018）提出分层时空算子（hierarchical spatio-temporal local binary pattern with integral projection，HSTLBP-IP），自适应分割空间特征块以提高分类能力。传统方法在面对微表情中微小的时空变化时，反应不够灵敏，导致识别准确率不高。

由于光流能够检测微小动作变化，一些工作通过光流法来提高微表情特征提取的精度。Liu等人（2016）提出的主方向平均光流（main directional mean optical-flow，MDMO）方法，结合局部运动信息和空间位置，降低特征维度并提高鲁棒性。Xu等人（2017）的面部动态图（facial dynamic map，FDM）方法利用光流场进行分块处理，通过主光流方向构建特征向量。Liu等人（2021）提出稀疏方向平均光流（sparse mean directional optical flow，SparseMDMO）结合图正则化稀疏编码去除冗余特征。但这些方法提取的大多是描述表情的基本、直观的特征和信息，并不涉及对脸部表情的深层次理解，计算复杂度较高，需要大量计算资源。

为了获取更丰富的特征，当前许多方法采用基于深度学习的特征提取技术。如，Xia等人（2020）使用循环卷积神经网络（recurrent convolutional neural network，RCN）分别提取原始图像和光流特征；Khor等（2018）将卷积神经网络（convolutional neural network，CNN）与长短期记忆网络（long short-term memory，LSTM）结合进行微表情识别。这些深度学习模型结构复杂、计算资源需求高，且在训练时容易出现过拟合。而其他方法，如浅层卷积网络（Gan等，2019；Khor等，2019），虽然简单易懂，但难以捕捉复杂特征。Ji等人（2013）提出的3D卷积神经网络（3DCNN）能同时提取时空特征，构建复杂的时空特征表达。然而，3DCNN在微表情数据集上容易过拟合，尤其在样本量较小的情况下。由于微表情的识别界限不明确且数据集不足，深度学习方法，包括3DCNN，面临特征学习能力和模型泛化能力的挑战。Gan等人（2019）提出结合光流特征与CNN的Apex框架（optical flow feature-apex network，OFF-ApexNet）以增强特征提取。Khor等人（2019）使用轻量级双流浅层网络来识别微表情，通过激活热图强调重要面部区域。Lei等人（2020）提出结合学习型视频运动放大与图卷积时序网络（learning-based video motion magnification and graph-based temporal convolutional network，Graph-TCN），通过放大局部肌肉运动特征提高识别效果。Van Quang等人（2019）将胶囊网络引入微表情识别，设计了高效的CapsuleNet。Liong等人（2019）提出浅层三流3DCNN（shallow triple stream three-dimensional convolutional neural network，STSTNet），通过轻量级计算提取高级特征并表达时空信息。Zhou等人（2022）提出基于双向Transformer的微表情检测网络，利用时空特征提取精确顶点帧检测。Lei等人（2021）通过深度卷积和Transformer编码器增强面部特征表示，提升识别效果。Wei等人（2022）提出注意力放大自适应网络（attention-based magnification adaptive network，AMAN），实现对不同微表情的自适应聚焦。Zhao等人（2022）提出深度原型学习框架ME-PLAN，通过知识转移与情景训练学习精确的ME特征。Zhai等人

(2023)利用自监督卷积位移模块与3层Transformer融合机制提取和融合动态特征。赵明华等人(2024)提出注意力引导的三流卷积神经网络(attention-guided three-stream convolutional neural network,ATSCNN),通过聚焦重要信息并抑制不相关信息来提高准确度。基于深度学习的方法提高了微表情识别的精度,但公开可用的微表情数据集的稀缺和训练数据的缺乏阻碍了这些方法学习有利于识别微表情的低级特征。此外,微表情本质上是复杂的,这些方法定义的分类边界往往不明确。

针对目前基于深度学习的微表情识别方法缺乏训练数据、无法学习到多类型的特征和识别精度低的问题,本文提出一种结合人脸图像特征、光流特征和光应变特征的微表情自动识别方法。主要贡献如下:1)提出一个由深度卷积神经网络与浅层3DCNN组成的三分支的表情自动识别模型,3个分支网络分别用于处理人脸特征、光流特征和光应变特征,并对3类特征进行逐层融合。这种多特征融合的方式充分利用了各类特征中隐含的不同信息,从而有效减少了识别过程中人脸细微特征的丢失,提高了微表情识别的精确度。2)提出和设计了多个注意力模块。提出面部特征增强注意力模块,加入在浅层3DCNN中,可使网络更专注于捕捉面部表情的微小变化;利用深度卷积神经网络ResNet-18进行迁移学习时,引入深度增强通道注意力网络ECANet来提取人脸特征,可使网络更有效地选择和关注特征通道,从而更好地捕捉关键特征并降低冗余信息的干扰。3)首次将IM Loss函数引入微表情识别的网络模型中,并结合Focal Loss函数,基于同一微表情类别内部不同微表情实例之间的相似性和差异性程度对样本进行加权,使得模型可以更好地处理难样本,有效提高了模型对于难样本的识别能力。4)大量实验表明,本文方法对微表情识别的准确率优于已有的微表情自动识别方法。

1 本文方法

本文提出一种基于多类型特征的微表情自动识别方法。本文网络模型是由1个深度卷积神经网络与2个浅层3DCNN构成的三分支结构,分别用于处理人脸特征、光流特征和光应变特征。其中,3DCNN具有同时提取时间和空间域特征的功能,可以有效提取出光流特征和光应变特征中包含的运动、形状变化和时间序列等重要的时空特性信息。对多类型特征的学习和融合可充分利用各类特征中隐含的不同信息,减少识别过程中人脸细微特征的丢失,有效提高微表情识别的精确度。在3个分支网络中分别提出和引入了新的注意力模块,能够使网络更好地捕捉关键特征和表情的微小变化,并降低冗余信息的干扰。本文首次将IM Loss函数引入用于微表情识别任务的网络模型中,有效提高了模型对于难样本的识别能力。本文方法还引入了中间帧增强技术,解决了训练样本缺乏的问题。

本文方法的流程分为预处理、特征提取和融合以及微表情分类3个步骤。在预处理阶段,首先通过顶点帧定位获取视频序列中的顶点帧位置。然后,利用顶点帧和起始帧提取光流信息,包括水平光流图、垂直光流图和光应变图。在特征提取和融合阶段,第1个分支网络引入ResNet-18(residual neural network 18)作为迁移模型,并加入深度增强通道注意力模块ECANet(efficient channel attention network)用于提取人脸特征。第2个和第3个分支网络均采用训练好的浅层3DCNN模型,并结合新的面部增强注意力模块FEAM(face-enhanced attention module),用于提取光流图的特征和光应变图的特征。将3个分支提取的特征进行逐层融合,以增强特征表达。在微表情分类阶段,使用归一化指数函数(softmax)对特征进行预测,得到分别属于Negative、Positive和Surprise 3个微表情类别的概率值,概率值最高的类别即为微表情的分类结果。本文方法利用多个分支对多类型特征进行提取和融合,并引入注意力机制,可以更好地捕捉微表情中的信息,从而提高微表情识别的性能。具体流程如图1所示。

1.1 预处理

1.1.1 顶点帧定位

受Gan等人(2019)的启发,本文从每个微表情序列中选择起始帧和顶点帧。在微表情序列中,顶点帧是指微表情的强度或表现最为明显的那一帧,而顶点帧定位是指确定微表情序列中顶点帧的位置。Liong等人(2018)表明,顶点帧中蕴含的信息能够在很大程度上表示整个微表情样本的特征。确定了顶点帧的位置后,就可以提取该帧前后的时间信息,进一步分析微表情的时序特性。本文实验采用的数据集CASME Ⅱ(Chinese Academy of Sciences micro-

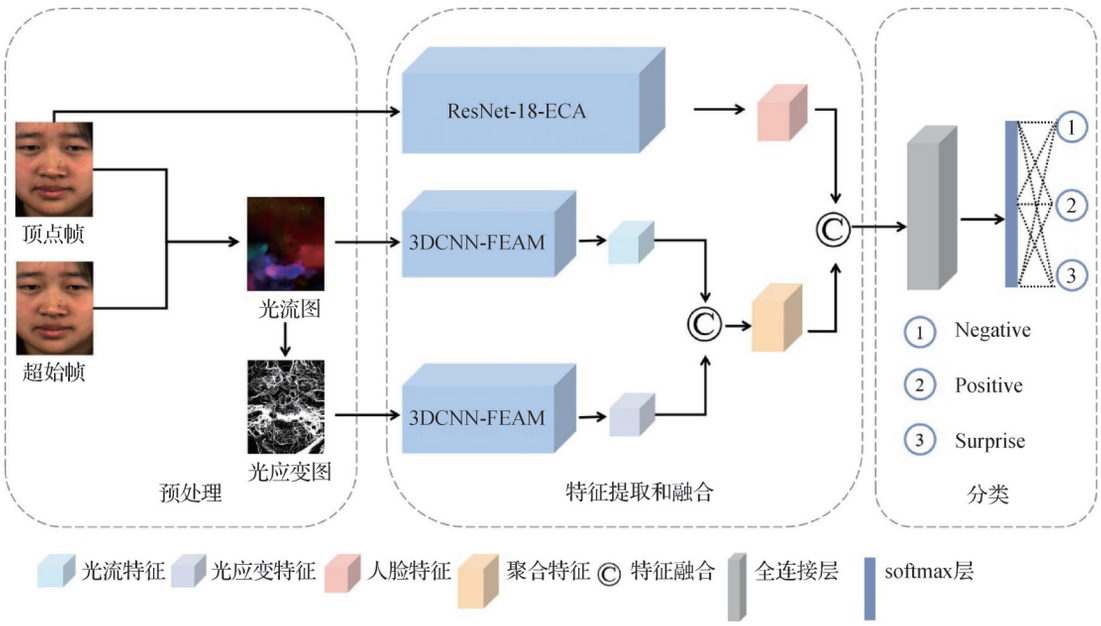

图1　本文方法流程图

Fig. 1　Flowchart of the method in this paper

expression dutabase Ⅱ)、SAMM(spontaneous actions and micro-morements)和MEGC2019(micro-expression grand challenge 2019 composite database)已经标记了起始帧和顶点帧。根据微表情的持续性和已有研究(Peng等,2019)可得:1)已标记的数据集显示,微表情的顶点帧通常出现在视频的中间部分;2)对于微表情的顶点帧前后的帧序列,其表情变化相对微小。基于此,对于未标记的SMIC(spontaneous micro-expression corpus)数据集,本文选取样本视频中间的那一帧作为顶点帧。本文使用顶点帧作为ResNet网络的输入来提取人脸特征,使用起始帧和顶点帧获取表示微表情动态变化的光流图像和光应变图像,将其作为3DCNN网络的输入以提取光流特征和光应变特征。

1.1.2　光流和光应变计算

光流法是一种对视频中的物体进行运动估计的方法。它是基于两帧图像之间局部导数的图像模式的近似,能够显示帧与帧之间明显的运动变化。通过光流场可以揭示视频中物体在连续两帧中的运动情况,光流场能够描述图像中每个像素的运动方向和幅度,更加明显地显示微表情在面部区域的变化,这使得模型能够学习到与微表情细微动作和变形相关的更加复杂和抽象的特征。另外,光流场还有助于描绘面部肌肉变形的细微程度,特别是对于微小的面部像素运动。

对于一个微表情序列$\boldsymbol{x}_t$,令$\boldsymbol{I}(x,y,t)$表示t时刻图像在像素点(x,y)处的亮度值。设该像素点在时间$\mathrm{d}t$内的移动距离为$(\mathrm{d}x,\mathrm{d}y)$,基于光流法亮度恒定原则,即同一物体相邻2帧的亮度不变,可以认为移动前后该像素的亮度值不变,即$I(x,y,t)=I(x+\mathrm{d}x,y+\mathrm{d}y,t+\mathrm{d}t)$。令$\boldsymbol{u}(x,y)=\frac{\mathrm{d}x}{\mathrm{d}t}$,$\boldsymbol{v}(x,y)=\frac{\mathrm{d}y}{\mathrm{d}t}$,使用起始帧和顶点帧计算光流引导特征,由这两帧计算得到的光流场可以表示为一个元组。即

$$\boldsymbol{O}_i=\{(\boldsymbol{u}(x,y),\boldsymbol{v}(x,y))\mid x=1,2,\cdots,X,\ y=1,\cdots,Y\} \tag{1}$$

式中,X和Y分别表示帧框架$f_{i,j}$的宽度和高度,$\boldsymbol{u}(x,y)$和$\boldsymbol{v}(x,y)$表示$\boldsymbol{O}_i$的水平和垂直分量。

由于应变模式仅与面部变形相关,且不易受照明条件和面部障碍物等因素影响,因此适用于微表情识别任务。通过处理光流矢量,可以获取光应变数据来描述面部运动模式,从而量化像素微小运动下的肌肉变形情况。光应变表示为

$$\boldsymbol{\varepsilon}=\frac{1}{2}[\nabla\boldsymbol{u}+(\nabla\boldsymbol{u})^{\mathrm{T}}] \tag{2}$$

式中,$\boldsymbol{u}=[\boldsymbol{u},\boldsymbol{v}]^{\mathrm{T}}$是位移矢量,表示三维空间中面部表情形变导致的位移在二维图像上的投影;∇表示对$\boldsymbol{u}$进行求导。

将式(2)展开为矩阵形式,具体为

$$\boldsymbol{\varepsilon}=\begin{bmatrix}\boldsymbol{\varepsilon}_{xx}=\dfrac{\partial\boldsymbol{u}}{\partial x} & \boldsymbol{\varepsilon}_{xy}=\dfrac{1}{2}\left(\dfrac{\partial\boldsymbol{u}}{\partial y}+\dfrac{\partial\boldsymbol{v}}{\partial x}\right)\\ \boldsymbol{\varepsilon}_{yx}=\dfrac{1}{2}\left(\dfrac{\partial\boldsymbol{v}}{\partial x}+\dfrac{\partial\boldsymbol{u}}{\partial y}\right) & \boldsymbol{\varepsilon}_{yy}=\dfrac{\partial\boldsymbol{v}}{\partial y}\end{bmatrix} \tag{3}$$

式中，对角应变分量(ε_{xx},ε_{yy})是法向应变分量，(ε_{xy},ε_{yx})是切应变分量。具体说来，法向应变衡量沿特定方向(如x轴或y轴)的长度变化，而切应变测量由剪切力引起的平面内角度变化。由于微表情中的肌肉运动可能在多个方向上产生影响，因此光应变可表示为法向应变和切应变分量的平方和，具体为

$$\left|\varepsilon_{x,y}\right| = \sqrt{\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + \varepsilon_{xy}^2 + \varepsilon_{yx}^2} \tag{4}$$

1.1.3　中间帧增强

为提高对微表情特征的学习精度，本文除了将顶点帧作为模型的输入以获取各类特征外，还纳入了微表情序列中的其他帧作为输入。基于微表情的瞬时特性，本文着重关注微表情顶点帧(即表情强度最大的帧)及其前后几帧。Liong 等人(2018)研究发现，顶点帧及其相邻帧在表情强度上的变化非常细微。通过观察与分析各类微表情序列，本文也发现顶点帧及其前后的帧表情差异很小。基于此，本文构建一个由顶点帧及其前后帧组成的中间帧序列，将该中间帧序列作为模型的输入，以丰富样本数据。对于顶点帧前后的帧数选择，本文在多个数据集上分别取顶点帧及前后 0 帧、3 帧、5 帧、7 帧、10 帧和 15 帧在同一条件下进行微表情识别实验，并计算未加权平均召回率(unweighted average recall，UAR)(计算过程详见 2.4 节)作为评估标准，表 1 展示了在 CASME Ⅱ和 SAMM 两个数据集上的实验结果。

从表 1 可见，与选取顶点帧前后 0 帧(即只选取顶点帧)相比，选取顶点帧及前后帧组成中间帧序列作为模型的输入可以提高微表情识别实验的 UAR 结果，说明纳入顶点帧的前后帧对提高识别精度是有效的。并且在 CASME Ⅱ数据集上选取顶点帧及前后 3 帧得到的 UAR 结果最优、时间效率最高，在 SAMM 数据集上选取顶点帧及前后 5 帧得到的 UAR 结果虽然最优，但与选取顶点帧及前后 3 帧得到的 UAR 结果相差很小，且时间效率大幅度低于选取顶点帧及前后 3 帧的时间效率。因此，综合考虑评估结果和时间效率，选取顶点帧及前后 3 帧的效果要更好。

表 1　CASME Ⅱ和 SAMM 数据集中选取不同中间帧序列的实验结果对比

Table 1　Comparison of experimental results for different intermediate frame sequences selected from CASME Ⅱ and SAMM datasets

帧数	CASME Ⅱ		SAMM	
	UAR	时间(测试)/min	UAR	时间(测试)/min
0	0.833 2	15.73	0.650 5	17.03
3	0.985 7	59.10	0.726 3	77.00
5	0.901 6	92.71	0.730 4	145.30
7	0.962 7	128.43	0.688 5	214.00
10	0.866 7	196.30	0.638 9	286.52
15	0.811 2	261.54	0.614 6	352.36

综上，本文构建了一个由顶点帧及其前后 3 帧组成的中间帧序列，并将这个中间帧序列作为模型的输入，实现了中间帧增强。图 2 展示了对 CASME Ⅱ数据集中的样本构建的中间帧序列。中间帧的使用可以增加数据的多样性，帮助网络模型更好地学习微表情的特征，因此能够获得更好的检测效果。

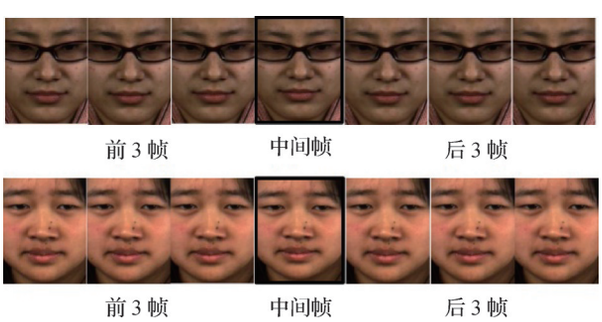

图 2　CASME Ⅱ数据集中的样本构建的中间帧序列

Fig. 2　Sequence of intermediate frames constructed from samples in the CASME Ⅱ dataset

1.2　特征提取和融合

1.2.1　光流特征和光应变特征提取

3D 卷积神经网络(3DCNN)中的卷积层都是三维的，由使用具有不同参数的 3D 卷积核来提取不同时间域中的时空特征，同时提取时间和空间域的特征。在光流和光应变特征中，包含了关于运动、形状变化和时间序列的重要信息，显然利用 3DCNN 网络有助于提取到更多关于这些时空特性的信息。因此，本文模型的第 2、3 个分支均引入 3DCNN 网络，分别用于提取光流特征和光应变特征。

3DCNN 网络的每个三维卷积层都有能力学习不同级别的特征，从低级特征如边缘和纹理到高级特征，使得网络能够逐层构建和表达复杂的特征层次结构。此外，3DCNN 网络具有较少的参数和计算要求，可以通过轻量级计算提取判别性高级特征和详细的微表情。由于微表情运动仅在特定的面部区

域内发生，大部分面部区域缺乏有效的分类特征，只有少数显示微表情运动的区域可以提供有用的信息来进行微表情分类。注意力机制可以帮助集中关注这些特定的面部区域，从而学习和捕捉关键特征。因此在使用3DCNN学习光流特征和光应变特征的基础上，本文提出并加入新的FEAM面部特征增强注意力模块，构建了新的分支网络3DCNN-FEAM，其网络配置和网络结构如图3所示。

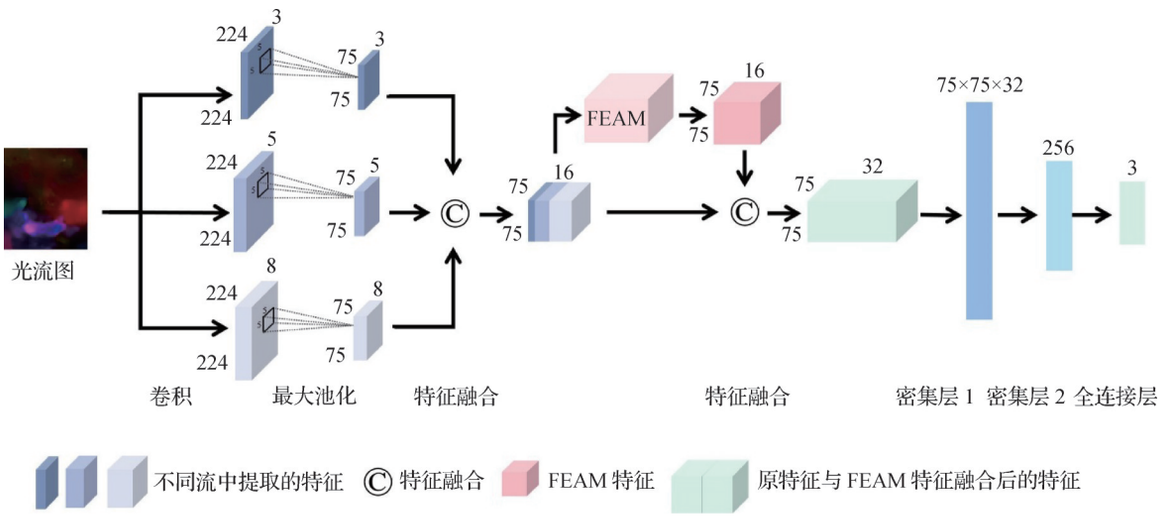

图3　3DCNN-FEAM网络

Fig. 3　3DCNN-FEAM network

将以上描述的光流图像和光应变图像分别进行重采样之后分别提取光流特征和光应变特征。在每个分支网络中，图像首先经过3个并行流，每个并行流都由一个卷积层和一个最大池化层组成。每个并行流使用不同参数的卷积核，提取不同大小的特征，通过不同数量的5×5卷积核补充小规模输入，以避免数据欠拟合的问题，然后将特征图按通道堆叠并进入最大池化层。利用最大池化操作可在消除冗余的同时突出主要特征。经过并行流之后，3个流中的特征在通道层面拼接在一起，以融合多尺度信息。融合后的特征通过FEAM面部特征增强注意力模块对特征进行组合和加权平均，以增强模型对关键信息的捕捉能力。然后，将经过FEAM面部特征增强注意力模块得到的特征与原特征进一步融合。最后，将融合后的特征经过3个全连接层。其中，前两个全连接层（也称为两个不同维度的密集层）分别将融合后的特征映射到更高级别的抽象表示，并逐步降低特征维度。第3个全连接层则将特征转换为适合该网络输出的维度，以支持特定任务的执行。

为了增强神经网络对面部特征的学习和提取能力，本文提出FEAM面部特征增强注意力模块，其结构如图4所示。卷积注意力模块在处理复杂或多样化的图像时有显著的适应性和稳定性，对捕捉快速且微弱的微表情极为有效，模块的高计算效率适用于资源受限的环境，与ECANet模块有明显的协同增效，进一步优化整体的识别性能。FEAM面部特征增强注意力模块在卷积注意力模块结构的基础上，引入双池化操作和全连接层，进一步提高特征图的表达能力，使神经网络更专注于捕捉面部的微小变化。该模块使得神经网络在学习过程中更加强调面部特征的提取，有助于更准确地捕捉面部表情、形态等关键信息。具体来说，本文引入了双池化操作，包括平均池化和最大池化，平均池化捕捉特征的整体信息，而最大池化则聚焦于特征的最显著部分。双池化操作使得网络在处理特征时，能够同时考虑不同尺度的信息，从而更全面地理解面部的结构。而全连接层的处理则有助于进一步整合特征信息，使得网络在学习过程中能够更好地理解不同特征之间的关系，提高了对面部细节的感知能力。

在FEAM结构中，本文将输入特征与注意力输出特征以相加的方式进行融合，通过这种融合方式，网络能够有效地提取深层信息，并将重点放在与面部相关的关键特征上。这种聚焦的处理方式有助于进一步增强网络在面部表情、姿态等细微变化的学习能力。

设输入到FEAM模块中的特征为$\boldsymbol{F}$，则$\boldsymbol{F}_{\max}$表示对输入特征$\boldsymbol{F}$应用最大池化操作得到的特征，$\boldsymbol{F}_{\text{avg}}$表示对输入特征$\boldsymbol{F}$应用平均池化操作得到的特征。$\boldsymbol{F}_{\text{fully}}$表示将$\boldsymbol{F}_{\max}$和$\boldsymbol{F}_{\text{avg}}$的合并结果输入到全连接层后得到的结果，表示为

$$\boldsymbol{F}_{\text{fully}} = \text{ReLU}(\boldsymbol{W}_1 \times (\boldsymbol{F}_{\max} \oplus \boldsymbol{F}_{\text{avg}})) \tag{5}$$

则通道注意力模块的输出$\boldsymbol{M}_{\text{chan}}$可以表示为

$$\boldsymbol{M}_{\text{chan}} = \sigma(\boldsymbol{W}_2 \times \boldsymbol{F}_{\text{fully}}) \tag{6}$$

式中，$\boldsymbol{W}_1$和$\boldsymbol{W}_2$分别表示全连接层的权重矩阵、通道注意力模块中全连接层的权重矩阵。

将$\boldsymbol{M}_{\text{chan}}$与通过全连接层输出的特征$\boldsymbol{F}_{\text{fully}}$进行通道注意力加权得到$\boldsymbol{F}_{\text{chan}}$，表示为

$$\boldsymbol{F}_{\text{chan}} = \boldsymbol{F}_{\text{fully}} \otimes \boldsymbol{M}_{\text{chan}} \tag{7}$$

空间注意力模块采用$\boldsymbol{F}_{\text{chan}}$的特征图作为输入，操作与通道注意力模块类似，但关注于空间维度，其输出$\boldsymbol{M}_{\text{spat}}$可表示为

$$\boldsymbol{M}_{\text{spat}} = \sigma(f_{\text{conv}}(\boldsymbol{F}_{\text{chan}})) \tag{8}$$

式中，f_{conv}为卷积操作。

将空间注意力特征$\boldsymbol{M}_{\text{spat}}$应用到经过通道注意力加权的特征$\boldsymbol{F}_{\text{chan}}$上，再次进行元素乘法（即空间注意力加权），可得

$$\boldsymbol{F}_{\text{spat}} = \boldsymbol{F}_{\text{chan}} \otimes \boldsymbol{M}_{\text{spat}} \tag{9}$$

连接$\boldsymbol{F}_{\text{spat}}$和原始输入特征$\boldsymbol{F}$，可得到最终的输出特征$\boldsymbol{F}_{\text{out}}$，表示为

$$\boldsymbol{F}_{\text{out}} = concat(\boldsymbol{F}, \boldsymbol{F}_{\text{spat}}) \tag{10}$$

式中，$concat$为特征融合操作。

通过在3DCNN中引入面部特征增强注意力模块，有效提高了特征图的表征能力，使得神经网络更加专注于面部细节的学习，同时通过输入特征层与注意力输出特征层的相加融合，很好地捕获了深层信息，从而更好地完成对面部信息的抽取与理解。

1.2.2　人脸特征提取

本文采用ResNet-18-ECA网络，即，将高效通道注意模块ECANet（efficient channel attention network）与ResNet-18组合的网络，提取微表情中间帧序列的人脸特征。该网络能够有效地提取和捕捉微表情图像中的关键信息，提高模型的性能和鲁棒性。ResNet-18作为经典的深度卷积神经网络，适用于图像特征提取。它具有多层卷积和池化层，能够自动学习和捕捉不同层次的特征，包括低级的纹理和高级的语义信息。而微表情往往包含微小而重要的面部动态变化。ECANet注意力机制与面部增强注意力模块FEAM具有明显的协同增效，进一步优化识别性能。ECANet注意力机制的引入可以进一步提升特征表示能力。

ECANet注意力机制专注于增强网络对特定通道的关注，通过自适应地调整通道的权重，可以突出那些包含微表情信息的通道，减少对不相关信息的关注。这种关注机制有助于提高网络的感知力和表达力。本文将ResNet-18与ECANet注意力机制结合在一起，能够在微表情识别任务中实现更深层次的特征提取和表示。ResNet-18提供了基础的特征提取能力，而ECANet注意力机制则进一步优化和加强了这些特征。这种协同作用使模型能够更好地捕捉微表情中微妙但重要的特征，提高了微表情识别的准确性和鲁棒性。

ECANet（enhanced channel attention network）是一个深度增强通道注意力模块，通过全局平均池化（global average pooling，GAP）和自适应最大池化（adaptive max pooling，AMP）计算通道注意力，进而

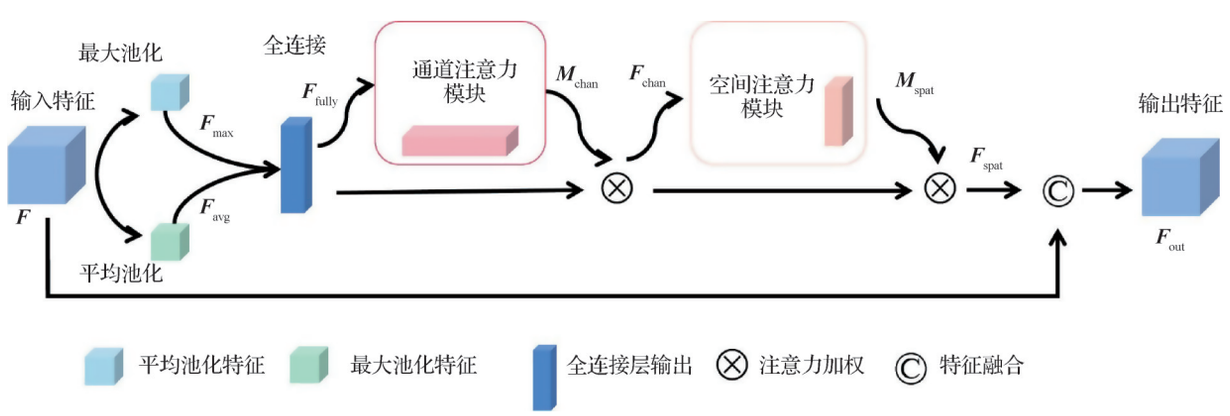

图4　FEAM面部增强注意力模块

Fig. 4　FEAM facial enhanced attention module

增强输入特征的表示能力。ECANet模块采用不降维的局部跨通道交互策略，可避免降维操作对通道注意力预测的不良影响，从而在有效降低模型复杂性的同时保持良好的性能。本文使用的ECANet模块具体结构如图5所示。ECANet模块包括全局平均池化操作（GAP）和自适应最大池化操作（AMP），用于捕获输入张量的关键特征。设人脸图像经过ResNet-18网络后输出的特征为$\boldsymbol{X}$，将$\boldsymbol{X}$作为该模块的输入，对其应用全局平均池化，即对维度为$H \times W \times C$的特征向量$\boldsymbol{X}$在每个通道上进行平均池化，产生一个空间维度为$1 \times 1 \times C$的特征向量，其中C是通道数，然后通过卷积和sigmoid激活处理得到通道注意力权重λ_1，将权重λ_1与原始输入$\boldsymbol{X}$相乘，得到输出特征$\tilde{\boldsymbol{\chi}}_1$。

自适应最大池化采用相似的步骤，对输入的特征$\boldsymbol{X}$应用自适应最大池化，通过卷积和sigmoid激活处理得到通道注意力权重λ_2，将权重与原始输入$\boldsymbol{X}$相乘，得到输出特征$\tilde{\boldsymbol{\chi}}_2$。两个权重分别与原始输入相乘，突出重要特征。最后将这些加权特征进行特征融合，形成最终输出。ECANet模块能够有效提高模型对不同特征的感知和性能，使网络更加高效地处理多尺度和多层次的信息。

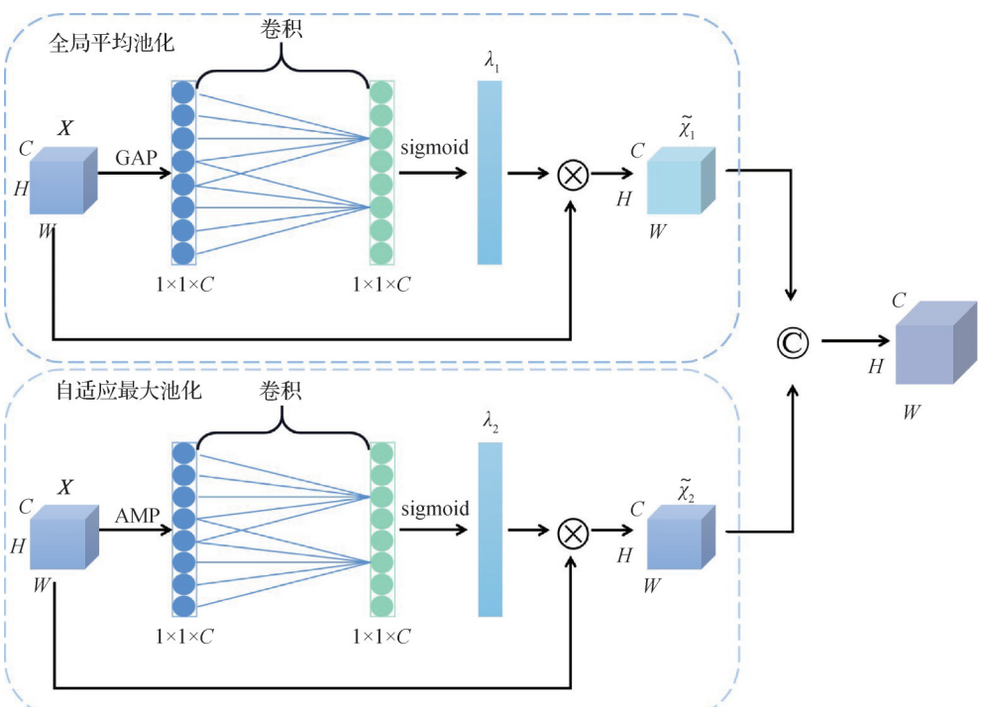

图5　ECANet模块

Fig. 5　ECANet module

1.2.3　特征融合

本文对3个分支网络得到的多类型特征进行逐层融合。具体地，首先将两个3DCNN-FEAM分支网络输出的光流特征和光应变特征进行融合合并，形成新的光流特征块；然后对光流特征块和ResNet-18-ECA网络输出的人脸特征块分别进行L2正则化处理，以控制模型复杂性和减少过拟合风险，并通对权重参数进行约束，防止模型出现不必要的噪声和过拟合问题；最后，将经过正则化处理的光流特征块和人脸特征块进行特征融合，得到聚合特征，用于后续的特征分类。

1.3　特征分类

将特征融合后得到的聚合特征输入到全连接层中，将聚合特征转换为适合于网络末端的softmax回归层的维度。softmax回归层对每个样本的特征进行分类。当特征输入到softmax层时，它会预测并输出样本属于3个类别（Negative、Positive和Surprise）的概率。最后，将每个样本概率最高的类别确定为该样本的微表情分类结果。

1.4　损失函数

对于微表情数据集中的每一个样本（即，一个微表情序列），经过上述网络模型处理后都会得到一个

对应的概率分布，这个分布由网络最后一层 softmax 层输出的每个类别的概率构成。本文提出新的损失函数，用于约束网络模型的预测概率分布，以提高模型对微表情识别的准确性。本文首次将 IM Loss 函数引入微表情识别，并结合 Focal Loss 函数共同约束模型，可有效解决微表情数据样本类别分布不平衡的问题。而微表情样本数据的缺乏是导致样本类别分布不平衡的重要原因之一，因此，提出的新损失函数可以较好地改善样本数据缺乏造成的样本分布不平衡，进而导致模型识别精度低的问题。其中，IM Loss 函数约束的是样本对应的概率分布，其目标是最大化预测概率分布和真实标签分布之间的关联程度，即两个分布之间共享信息的量，从而促使模型生成的概率分布尽可能地接近真实的分布，继而增强模型对所有类别的识别能力。Focal Loss 函数约束的是特定类别的预测概率，即主要关注单个概率值，特别是那些难以预测或分类错误的样本的概率值，通过调整它们的损失权重可以提高模型对少数类的识别能力。由于概率分布取决于概率值，所以 Focal Loss 函数也可看做是对样本对应概率分布的约束。

1.4.1 IM Loss

IM Loss 函数常用于解决无监督域适应中目标域中无标签数据分类的问题，可在确保分类标签的多样性和平均性的同时增强目标域中不容易分类样本的置信度。本文首次将 IM Loss 函数引入微表情识别中，旨在增强模型对于微表情分类的置信度并确保各类微表情的平衡识别。IM Loss 函数由两部分组成，分别是熵最小化部分和多样性最大化部分。

熵最小化部分主要是为了确保模型输出的预测标签分布对于给定的输入样本是尽可能确定的，即，减少模型的不确定性。在数学上，熵是用来衡量一个随机变量不确定性的量度，对于一个概率分布，其熵越小意味着这个分布越倾向于某一类，不确定性越小。在分类任务中，熵最小化通常意味着模型对于其预测结果很有信心。

多样性最大化部分旨在确保模型对不同的输入样本给出不同类别的预测结果，即增加模型输出的多样性。这对于那些样本在不同域或样本类别不平衡的场景中的任务尤其重要，因为模型可能会对多数类别的样本产生过拟合，忽略少数类别。通过最大化多样性，可以确保每一个类别都得到适当的关注，从而使得整体的预测更加均衡。

IM Loss 函数的目的是尽可能让分类的标签具有多样性，即每个类别都比较平均。为了达到这一目的，只需计算出整个目标域输出的平均值，使其熵最大，让不易分类的样本远离分类界限，增加置信度。IM Loss 函数定义为

$$\begin{cases} L_{\mathrm{IM}}(p_{k(\boldsymbol{x}_t)}) = L_{\mathrm{ent}}(p_{k(\boldsymbol{x}_t)}) + L_{\mathrm{div}}(p_{k(\boldsymbol{x}_t)}) \\ L_{\mathrm{ent}}(p_{k(\boldsymbol{x}_t)}) = -\mathbb{E}_{\boldsymbol{x}_t \in \mathbf{R}} \sum_{t=1}^{N} \sum_{k=1}^{K} p_{k(\boldsymbol{x}_t)} \log(p_{k(\boldsymbol{x}_t)}) \\ L_{\mathrm{div}}(p_{k(\boldsymbol{x}_t)}) = \sum_{k=1}^{K} \hat{p}_k \log \dfrac{\hat{p}_k}{1/K},\ \hat{p}_k = \dfrac{1}{V} \sum_{t=1}^{N} p_{k(\boldsymbol{x}_t)} \end{cases} \tag{11}$$

式中，L_{ent} 表示 IM Loss 函数的熵最小化部分，用于计算每个样本预测概率的交叉熵损失，并在整个数据集上取平均。这里，设微表情数据集为 $\boldsymbol{R} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_V\}$，其中样本 $\boldsymbol{x}_t (t = 1, 2, \cdots, V)$ 表示微表情数据集 $\boldsymbol{R}$ 中的一个微表情序列，V 为数据集 $\boldsymbol{R}$ 中样本的总数。$\mathbb{E}_{\boldsymbol{x}_t \in \mathbf{R}}$ 表示对微表情数据集 $\boldsymbol{R}$ 中的每个样本 $\boldsymbol{x}_t$ 对应的损失加和求平均值，每个样本 $\boldsymbol{x}_t$ 对应的损失即在模型当前参数下该样本 $\boldsymbol{x}_t$ 所有可能类别的预测概率的交叉熵损失的总和。K 为数据集中样本类别的总数，$p_{k(\boldsymbol{x}_t)}$ 为模型预测样本 $\boldsymbol{x}_t$ 属于类别 k 的概率值。L_{div} 表示 IM Loss 函数的多样性最大化部分，用于直接计算每个样本 $\boldsymbol{x}_t$ 预测的概率分布与均匀分布之间的 KL（Kullback Leibler）散度，以此来促进预测的多样性，避免对某些类别的偏好，并对这些值求和。$\hat{p}_k$ 为每个样本被预测为类别 k 的平均概率，$1/K$ 是均匀分布的预期概率，本文假设每个类别选中的概率是相等的。

1.4.2 Focal Loss

Focal Loss 函数是为了解决分类任务中的类别不平衡问题而设计的损失函数。在传统的交叉熵损失中，所有样本均平等对待，这很容易导致模型被数量庞大且容易分类的样本所支配，从而忽视那些少数但是难以分类的样本。为了解决这个问题，Focal Loss 函数中引入了一个可调节的因子，这个因子可以动态调整每一个样本的权重。具体来说，该可调节的因子可降低易分类样本的权重，增加难以分类样本的权重，从而使模型更加关注那些难以分类的样本。这种设计有助于缓解因类别不平衡带来的问题，使得模型在面对各种类别的样本时都能有较好的分类效果。Focal Loss 函数定义为

$$L_{\text{focal}}(p_{k(x_t)}) = -\alpha_k(1 - p_{k(x_t)})^{\gamma}\log(p_{k(x_t)}) \quad (12)$$

式中，$p_{k(x_t)}$表示模型将x_t预测为类别k的概率值，Focal Loss针对数据集中的每个样本独立计算其损失。式中，$k \in [k_1, k_2, k_3]$，k_1为类别Negative，k_2为类别Positive，k_3为类别Surprise。α_k是预设的权重参数，用来平衡不同类别的权重，对于负样本该权重设置得较小。γ是调节参数，通过设置γ可以增加难以分类样本的权重，从而增加模型对难以分类样本的关注度。

1.4.3 Total Loss

为了提高模型对难分类样本的识别能力，并确保输出的多样性和确定性，本文采用基于样本类别分类的难易程度进行样本加权的策略。具体来说，Focal Loss通过动态调整权重来更加关注难分类的样本，以应对类别不平衡问题。而IM Loss函数则在提高模型确定性的同时增加模型输出的多样性。这种策略使得模型不仅能够在数据集的整体层面上对难样本给予足够关注，还能够在特定类别的内部区分难以辨别的样本，从而提高了模型的分类精确度和鲁棒性。综上，本文使用的Total Loss函数定义为

$$L_{\text{total}}(\boldsymbol{x}_t) = \alpha L_{\text{focal}}(p_{k(x_t)}) + \beta L_{\text{IM}}(p_{k(x_t)}) \quad (13)$$

式中，α和β是权重系数，用于调整两种损失的比重。

2 实 验

为了证明本文提出的结合多特征融合和注意力机制的微表情识别方法的有效性，首先，介绍所采用的数据集以及性能评估方法；接着，进行该方法与其他方法的对比分析；最后，通过专门的消融实验评估模型的稳定性。

2.1 微表情数据集与数据整理

本文实验采用CASME Ⅱ、SMIC、SAMM和MEGC2019四个数据集。1）中国科学院微表情数据集CASME Ⅱ（Yan等，2013）。CASME Ⅱ包含247个来自50名男性和50名女性的自发样本。2）芬兰奥卢大学的自发微表情数据集SMIC（Li等，2013）。该数据集通过100帧/s的摄像头采集了16名被试对象的164个微表情。所有类型的微表情被分为Negative、Positive和Surprise 3类。3）英国曼彻斯特大学的自发动作和微表情数据集SAMM（Davison等，2018），该数据集包含159个由200帧/s相机记录的微表情样本，样本来自32名参与者。这些参与者的被试平均年龄为33.24岁，性别比例相当。4）MEGC2019（See等，2019）为复合数据集，包含来自不同背景（种族、环境和性别的68名受试者的442个样本。根据第2届微表情识别大赛中使用的简化情感类别，按照SMIC数据集的情感分类（Negative、Positive、Surprise），MEGC2019复合数据集对CASME Ⅱ和SAMM数据集进行了重新分类，以统一样本标签并将现有标签重新映射到新的标签空间中。重新分类旨在减少由不同刺激和环境设置造成的情绪类别的模糊性。具体操作为：将“厌恶”、“愤怒”、“压抑”、“轻蔑”、“悲伤”和“恐惧”样本归为“Negative”样本；将“高兴”样本归为“Positive”样本；“惊讶”样本保持不变，仍为“Surprise”；标有“其他”的样本被排除在外，因为它们是不可分类的。MEGC2019数据集具体信息如表2所示。

2.2 实验环境和设置

本文实验采用的机器配置信息及开发运行环境为：Ubuntu 20.04；GeForce RTX 3090；CUDA11.0；PyTorch 1.7.1+cu110；Pycharm 2019.3.5；Python 3.7。本文使用Adam算法优化模型所有模块，初始学习速率设为0.000 5，对模型训练300个epochs，batch size为256，将权重参数α_k、γ、α和β分别设为0.5、2、0.5和0.5。

表2 MEGC2019数据集细节

Table 2 Details of the MEGC2019

标签类别	CASME Ⅱ		SMIC		SAMM		合计
	情绪标签	数量	情绪标签	数量	情绪标签	数量	
Negative	厌恶(27)压抑(61)	88	消极	70	愤怒(57)蔑视(12)厌恶(9)恐惧(8)悲伤(6)	92	250
Positive	快乐	32	积极	51	快乐	26	109
Surprise	惊讶	25	惊讶	43	惊讶	15	83
合计		145		164		133	442

2.3 实验细节

为了减少网络模型训练过程中的随机性和潜在偏差，本文使用留一交叉验证法（leave-one-subject-out，LOSO）进行交叉验证。该方法将数据集中一个受试者的所有微表情数据都保留用于测试，其余数据用于训练。确保在训练阶段排除测试对象的信息。该过程迭代重复 y 次，y 对应微表情数据集中的受试者数量。SMIC、CASME Ⅱ和 SAMM 数据集分别包含 16、24 和 28 位受试者，因此需要执行 68 轮 LOSO，即 $y = 68$。

2.4 实验评估标准

本文实验使用的评估标准包括未加权平均召回率（unweighted average recall，UAR）和未加权 F1 得分（unweighted F1-score，UF1）。

*UF*1 即宏观平均 *F*1 分数，通过平均所有类别的 *F*1 分数来计算的 *UF*1 在多类问题中特别有用，因为它对频率较低的类具有同等重要性。为了计算 *UF*1，首先确定每个类别 k 的 *F*1 值，然后对这些分数进行平均以产生 *UF*1 分数。

$$F1_k = \frac{2TP_k}{2TP_k + FP_k + FN_k} \tag{14}$$

$$UF1 = \frac{1}{Q}\sum_{k=1}^{N} F1_k \tag{15}$$

式中，Q 为微表情的标签数量，$F1_k$ 为类别 k 的分类结果的 *F*1 值，TP_k、FP_k 和 FN_k 分别代表类别 k 的分类结果中真阳性、假阳性和假阴性的样本数量。

未加权平均召回率（*UAR*）也称平衡准确率，通过将每个类别的准确率相加后对类别数量求平均值来获得。

$$UAR = \frac{1}{Q}\sum_{c=1}^{N}\frac{TP_c}{M_c} \tag{16}$$

MEGC2019 复合数据集中不同类别的样本数量是不平衡的，Surprise、Positive 和 Negative 3 类表情样本的比例大约为 1∶1.3∶3，存在严重的数据类别不平衡问题。因此，本文使用 *UAR* 和 *UF*1 指标评估所提出方法进行三分类识别的性能，这两个指标通过平等对待每个类，确保模型正确识别频率较低的类的能力也被考虑在内，从而更公平地评估其性能。

2.5 对比实验结果

2.5.1 三分类对比实验结果

在相同实验条件下，将本文方法与目前主流的三分类识别方法进行比较，包括基于手工提取特征的方法（Zhao 和 Pietikainen，2007；Liong 等，2018）和基于深度学习的方法（Gan 等，2019；Van Quang 等，2019；Liong 等，2019；Xia 等，2020；Lei 等，2021；Zhou 等，2022；Zhao 等，2022；Zhai 等，2023），在 4 个整理后的数据集 CASME Ⅱ、SMIC 和 SAMM 以及 MEGC2019 复合数据集上对方法进行评估。不同方法的 *UF*1 和 *UAR* 结果如表 3 所示。

表 3 不同方法在三分类数据集上的微表情识别性能

Table 3 Micro-expression recognition performance of different methods on triple categorical dataset

类别	方法	CASME Ⅱ		SMIC		SAMM		MEGC2019	
		*UF*1	*UAR*	*UF*1	*UAR*	*UF*1	*UAR*	*UF*1	*UAR*
传统方法	LBP-TOP（Zhao 和 Pietikainen，2007）	0.702 6	0.742 9	0.200 0	0.528 0	0.395 4	0.410 2	0.588 2	0.578 5
	Bi-WOOF（Liong 等，2018）	0.780 5	0.802 6	0.572 7	0.582 9	0.521 1	0.513 9	0.629 6	0.622 7
深度学习方法	OFF-ApexNet（Gan 等，2019）	0.876 4	0.868 1	0.681 7	0.669 5	0.540 9	0.539 2	0.719 6	0.709 6
	CapsuleNet（Van Quang 等，2019）	0.706 8	0.701 8	0.582 0	0.587 7	0.588 2	0.598 9	0.652 0	0.650 6
	STSTNet（Liong 等，2019）	0.838 2	0.868 6	0.680 1	0.701 3	0.658 8	0.681 0	0.735 3	0.760 5
	RCN（Xia 等，2020）	0.808 7	0.856 3	0.598 0	0.599 1	0.677 1	0.697 6	0.705 2	0.716 4
	AU-GCN（Lei 等，2021）	0.879 8	0.871 0	0.719 2	0.721 5	0.775 1	0.789 0	0.791 4	0.793 3
	FeatRef（Zhou 等，2022）	0.891 5	0.887 3	0.701 1	0.708 3	0.737 2	0.715 5	0.783 8	0.783 2
	ME-PLAN（Zhao 等，2022）	0.863 2	0.877 8	0.712 7	0.725 6	0.716 4	0.741 8	0.771 5	0.768 4
	FRL-DGT（Zhai 等，2023）	0.919 0	0.903 0	0.743 0	0.749 0	0.772 0	0.758 0	0.812 0	0.811 0
	本文	**0.984 3**	**0.985 7**	**0.767 1**	**0.752 5**	**0.787 1**	**0.790 4**	**0.822 1**	**0.830 7**

注：加粗字体表示各列最优结果。

从表3可以看出，基于深度学习的方法性能优于传统的基于手工特征提取的方法。在样本分类极度不平衡的MEGC2019复合数据集上，与基于手工特征的经典方法LBP-TOP相比，本文方法的*UF*1和*UAR*分别提高23.39%和25.22%；与同样使用顶点帧以及光流的方法STSTNet相比，本文方法在MEGC2019数据集上的*UF*1提高8.68%，*UAR*提高7.02%。

本文方法在CASME Ⅱ数据集上获得的*UF*1和*UAR*性能最好，分别为0.984 3和0.985 7，在SMIC数据集上获得的*UF*1和*UAR*结果也优于目前已有方法，分别为0.767 1和0.752 5。与基于手工特征的经典方法中最好的结果相比，本文方法的*UF*1和*UAR*在CASME Ⅱ数据集上分别提高20.38%和18.31%，在SMIC数据集上分别提高19.44%和16.96%，在SAMM数据上分别提高26.60%和27.65%。此外，与基于深度学习的方法中最好的结果相比，本文方法的*UF*1和*UAR*在CASME Ⅱ数据集上分别提高6.53%和8.27%，在SMIC数据集上分别提升2.41%和0.35%，在SAMM数据集上分别取得1.51%和3.24%的提升。图6为MEGC2019复合数据集在不同方法上的混淆矩阵，由图6可知，本文方法在3个类别上的结果基本优于其他方法。综上，可以充分说明本文提出的方法进行三分类微表情识别的有效性。

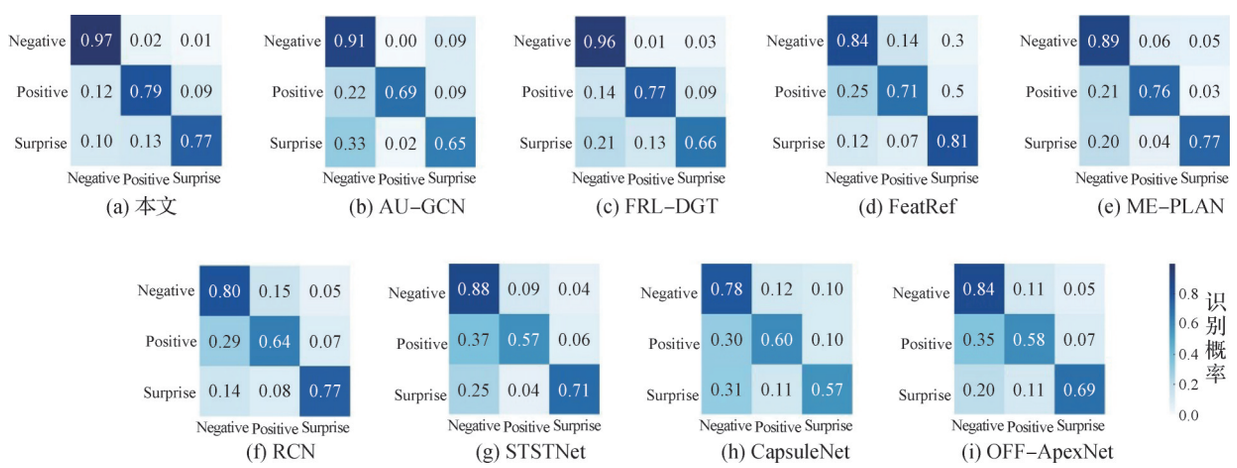

图6　MEGC2019数据集在不同方法上的混淆矩阵

Fig. 6　Confusion matrix for the MEGC2019 dataset on different methods ((a) ours; (b) AU-GCN; (c) FRL-DGT; (d) FeatRef; (e) ME-PLAN; (f) RCN; (g) STSTNet; (h) CapsuleNet; (i) OFF-ApexNet)

2.5.2　多分类对比实验结果

本文方法可以轻松扩展做多分类微表情识别，通过改变网络架构中每个分支网络末端的全连接层的维度和softmax回归层的维度即可实现多分类的微表情识别。由于SMIC数据集和MEGC2019数据集只标注了3类表情，故本文选择CASME Ⅱ和SAMM数据集进行五分类表情实验，这2个数据集均包括“厌恶”、“高兴”、“其他”、“压抑”和“惊讶”类微表情。

本文方法与现有的一些方法的比较结果如表4所示。可见，本文方法在CASME Ⅱ和SAMM数据集上的*UF*1和*UAR*结果明显高于其他方法，在CASME Ⅱ数据集上，本文方法比位于第2位的TSCNN方法的*UF*1提高了1.92%，*UAR*提高0.52%。本文方法的*UF*1和*UAR*在SAMM数据集上也分别比其他最优方法提高2.82%和2.29%。图7为CASME Ⅱ数据集在不同方法上的混淆矩阵，由图7可知，本文方法在多分类上的结果基本优于其他方法。

2.6　消融实验

2.6.1　网络架构消融

为验证本文模型采用多特征融合机制和多注意力机制的有效性，使用LOSO处理从视频序列中提取的图像帧，设计了7个消融实验。为叙述方便，将模型中用于提取人脸特征的第1个网络分支记为N_A，将用于提取光流特征的第2个网络分支记为N_B，将用于提取光应变特征的第3个网络分支记为N_C，将完整的网络模型记为N_D，将第1个网络分支中的深度增强通道注意力模块ECANet记为M_E，将第2个和第3个网络分支中的面部特征增强注意力模块

表 4　不同方法进行多分类微表情识别的识别率比较

Table 4　Comparison of the performace of different methods for multi-classfication micro-expression recognition

类别	方法	CASME Ⅱ		SAMM	
		*UF*1	*UAR*	*UF*1	*UAR*
传统方法	LBP-TOP(Zhao 和 Pietikainen，2007)	0.396 8	0.358 9	0.355 6	0.289 2
	Bi-WOOF(Liong 等，2018)	0.610 0	0.588 0	0.397 0	0.583 0
深度学习方法	AlexNet(Krizhevsky 等，2012)	0.629 6	0.667 5	0.426 0	0.529 4
	DSSN(Khor 等，2019)	0.729 7	0.707 8	0.464 4	0.573 5
	Graph-TCN(Lei 等，2021)	0.724 6	0.739 8	0.698 5	0.750 0
	AMAN(Wei 等，2022)	0.710 0	0.754 0	0.670 0	0.688 5
	TSCNN(Song 等，2019)	0.807 0	0.809 7	0.694 2	0.717 6
	本文	**0.826 2**	**0.814 9**	**0.722 4**	**0.740 5**

注：加粗字体表示各列最优结果。

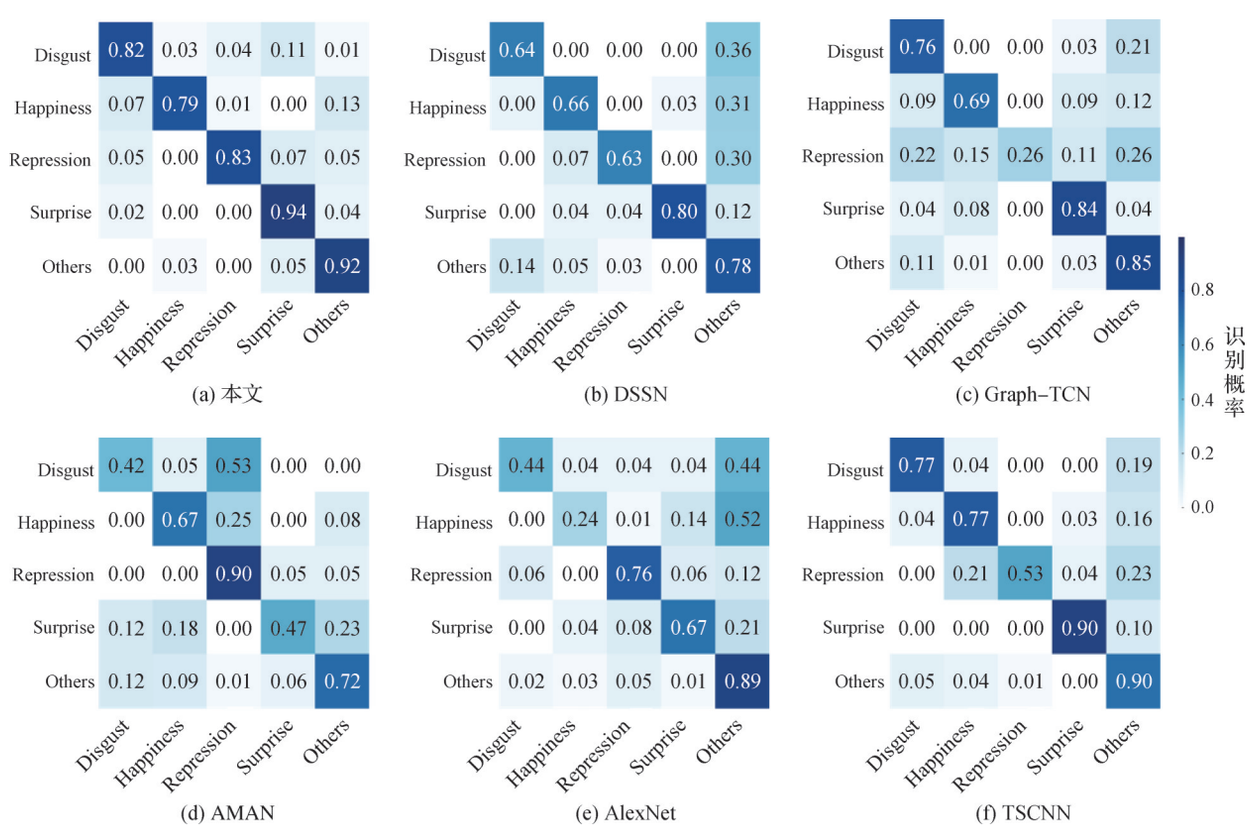

图 7　CASME Ⅱ数据集在不同方法上的混淆矩阵

Fig. 7　Confusion matrix for CASME Ⅱ dataset on different methods

((a) ours；(b) DSSN；(c) Graph-TCN；(d) AMAN；(e) AlexNet；(f) TSCNN)

FEAM 记为 M_F。

网络架构消融实验 1 为完整模型去掉第 1 个人脸特征提取分支(N_D–N_A)；架构实验 2 为完整模型去掉第 2 个光流特征提取分支(N_D–N_B)；架构实验 3 为完整模型去掉第 3 个光应变特征提取分支(N_D–N_C)；架构实验 4 为完整模型去掉 ECANet 模块(N_D–M_E)；架构实验 5 为完整模型的第 2 个分支去掉 FEAM 模块(N_B–M_F)；架构实验 6 为完整模型的第 3 个分支去掉 FEAM 模块(N_C–M_F)；架构实验 7 为完整模型(N_D)。

7 种消融方法在 CASME Ⅱ数据集上的对比结果如表 5 所示。可见，去掉任何一个特征提取分支或注意力模块，都会导致方法性能的下降，说明本文提出

的多特征融合机制和多注意力机制有效提升了模型的微表情识别精度。另外，表5还展示了各个架构消融实验中的网络参数量(Param)。可见，参数量主要集中在ResNet-18模块，该模块虽然参数量占比大，但对算法性能有重要影响，去掉该模块会导致识别精度显著下降。其他各模块均为轻量级的。本文完整模型能以较低的参数量达到较高的识别精度。

表5　网络架构消融实验结果

Table 5　Network architecture ablation experiment results

实验	*UF*1	*UAR*	Param/M
架构实验1 $N_D - N_A$	0.712 8	0.781 8	1.43
架构实验2 $N_D - N_B$	0.751 4	0.717 9	12.57
架构实验3 $N_D - N_C$	0.948 4	0.925 3	12.57
架构实验4 $N_D - M_E$	0.916 5	0.897 5	13.13
架构实验5 $N_B - M_F$	0.953 9	0.939 2	13.11
架构实验6 $N_C - M_F$	0.931 9	0.917 4	13.11
架构实验7 N_D(本文)	**0.984 3**	**0.985 7**	13.22

注：加粗字体表示各列最优结果。

2.6.2　损失函数消融

为验证本文提出的新损失函数的有效性，使用LOSO，设计了5个消融实验。由于本文方法中的Focal Loss函数，其参数γ设置为0时，即退化为现有的微表情识别方法中常用的交叉熵损失函数，因此损失函数消融实验1采用最基础的交叉熵损失函数，记为L_{CE}，作为baseline；损失实验2采用交叉熵损失函数L_{CE}和IM Loss函数L_{IM}；损失实验3仅采用Focal Loss函数L_{focal}；损失实验4仅采用IM Loss函数L_{IM}；损失实验5采用Focal Loss函数L_{focal}和IM Loss函数L_{IM}，即本文完整的损失函数L_{total}。

5种损失函数消融方法在CASME Ⅱ数据集上的对比实验结果如表6所示。可见，与使用本文完整损失函数L_{total}(损失实验5)相比，仅使用其中的Focal Loss函数L_{focal}(损失实验3)或IM Loss函数L_{IM}(损失实验4)作为约束项，得到的*UF*1和*UAR*指标均明显下降。另外，虽然仅使用Focal Loss函数L_{focal}(损失实验3)的*UF*1和*UAR*值均低于仅使用交叉熵损失函数L_{CE}(损失实验1)的*UF*1和*UAR*值，但联合使用Focal Loss函数和IM Loss函数(损失实验5)的*UF*1和*UAR*值均明显高于联合使用交叉熵损失函数和IMLoss函数(损失实验2)的*UF*1和*UAR*值，分别提升7.60%和8.17%。综上，证明本文提出的由Focal Loss函数和IM Loss函数构成的总损失函数效果最优。

表6　损失函数消融实验结果

Table 6　Loss function ablation experiment result

约束项	*UF*1	*UAR*
损失实验1 L_{CE}	0.889 9	0.889 7
损失实验2 $L_{CE} + L_{IM}$	0.908 3	0.904 0
损失实验3 L_{focal}	0.870 5	0.863 1
损失实验4 L_{IM}	0.866 3	0.851 2
损失实验5 $L_{total}(L_{focal} + L_{IM})$	**0.984 3**	**0.985 7**

注：加粗字体表示各列最优结果。

2.7　时间评估

为评估本文模型的运行效率，测试了其在4个数据集上的执行时间，如表7所示。本文采用LOSO方法进行运行效率的评估。该方法可严格可确保测试集中每个受试者的数据均被评估，因此需要多次训练和测试。相比其他评估方法，该方法在确保结果的可靠性的同时，也增加了执行时间。但整体来看，本文模型在不同数据集上均表现出较高的运行效率。

表7　本文方法在4个数据集上的执行时间

Table 7　Execution time of the method in this paper on 4 datasets

数据集	执行时间/s
CASME Ⅱ	14.561 1
SMIC	17.220 0
SAMM	19.470 3
MEGC2019	17.042 4

2.8　错例分析

图8展示了本文方法的错误分类案例。通过注意力图分析发现，模型在识别时对嘴角和眼部等关键区域关注不足，尤其在微表情细微变化时易发生误判，为模型优化提供了参考依据。

3　结　论

本文提出一种融合多类型特征和多注意力机制的多分支微表情自动识别网络模型。模型中并行的

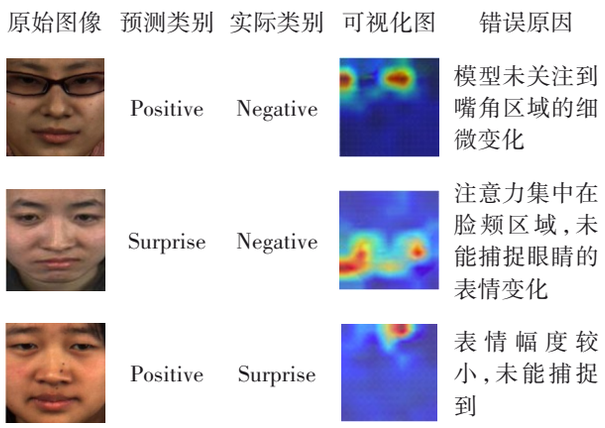

图 8　本文方法错误分类案例分析

Fig. 8　Case study of misclassification of proposed method

1 个深度卷积神经网络和 2 个浅层 3D 卷积神经网络分别用于人脸特征、光流特征和光应变特征的提取。这些不同类型特征的充分融合可有效减少人脸细微特征的损失，提升微表情识别的精确度。在浅层 3DCNN 中提出面部特征增强注意力模块，在深度卷积神经网络中引入深度增强通道注意力模块，这种多注意力机制可使模型更有针对性地捕捉人脸关键特征并降低冗余信息的干扰。构建的结合 IM Loss 约束项和 Focal Loss 约束项的新损失函数，显著提升了模型对于难样本的识别能力和识别精度。本文方法在微表情识别精度上取得了显著提升，但在模型的泛化能力上仍有一定不足。在处理来自真实世界不同场景下的微表情视频，或与训练集差异巨大的数据集时，识别精度会受到一定影响。未来工作中，拟结合迁移学习和预训练模型来解决这个问题。

参考文献（References）

Davison A K, Lansley C, Costen N, Tan K and Yap M H. 2018. SAMM: a spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing, 9(1): 116-129 [DOI: 10.1109/TAFFC.2016.2573832]

Ekman P. 2003. Darwin, deception, and facial expression. Annals of the New York Academy of Sciences, 1000(1): 205-221 [DOI: 10.1196/annals.1280.010]

Gan Y S, Liong S T, Yau W C, Huang Y C and Tan L K. 2019. OFF-ApexNet on micro-expression recognition system. Signal Processing: Image Communication, 74: 129-139 [DOI: 10.1016/j.image.2019.02.005]

Huang X H, Wang S J, Zhao G Y and Piteikainen M. 2015. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection//Proceedings of 2015 IEEE International Conference on Computer Vision Workshop. Santiago, Chile: IEEE: 1-9 [DOI: 10.1109/ICCVW.2015.10]

Huang X H, Zhao G Y, Hong X P, Zheng W M and Pietikäinen M. 2016. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing, 175: 564-578 [DOI: 10.1016/j.neucom.2015.10.096]

Ji S W, Xu W, Yang M and Yu K. 2013. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231 [DOI: 10.1109/TPAMI.2012.59]

Khor H Q, See J, Liong S T, Phan R C W and Lin W Y. 2019. Dual-stream shallow networks for facial micro-expression recognition//Proceedings of 2019 IEEE international conference on image processing (ICIP). Taipei, China: IEEE: 36-40 [DOI: 10.1109/ICIP.2019.8802965]

Khor H Q, See J, Phan R C W and Lin W Y. 2018. Enriched long-term recurrent convolutional network for facial micro-expression recognition//Proceedings of the 13th IEEE international conference on automatic face and gesture recognition (FG 2018). Xi'an, China: IEEE: 667-674 [DOI: 10.1109/FG.2018.00105]

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: 1097-1105

Lei L, Chen T, Li S G and Li J F. 2021. Micro-expression recognition based on facial graph representation learning and facial action unit fusion//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1571-1580 [DOI: 10.1109/CVPRW53098.2021.00173]

Lei L, Li J F, Chen T and Li S G. 2020. A novel graph-TCN with a graph structured representation for micro-expression recognition//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2237-2245 [DOI: 10.1145/3394171.3413714]

Li X B, Pfister T, Huang X H, Zhao G Y and Pietikäinen M. 2013. A spontaneous micro-expression database: inducement, collection and baseline//Proceedings of the 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (FG). Shanghai, China: IEEE: 1-6 [DOI: 10.1109/FG.2013.6553717]

Liong S T, Gan Y S, See J, Khor H Q and Huang Y C. 2019. Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition//Proceedings of the 14th IEEE international conference on automatic face and gesture recognition (FG 2019). Lille, France: IEEE: 1-5 [DOI: 10.1109/FG.2019.8756567]

Liong S T, See J, Wong K S and Phan R C W. 2018. Less is more: micro-expression recognition from video using apex frame. Signal Processing: Image Communication, 62: 82-92 [DOI: 10.1016/j.

image.2017.11.006]

Liu Y J, Li B J and Lai Y K. 2021. Sparse MDMO: learning a discriminative feature for micro-expression recognition. IEEE Transactions on Affective Computing, 12(1): 254-261 [DOI: 10.1109/TAFFC.2018.2854166]

Liu Y J, Zhang J K, Yan W J, Wang S J, Zhao G Y and Fu X L. 2016. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Transactions on Affective Computing, 7(4): 299-310 [DOI: 10.1109/TAFFC.2015.2485205]

Peng M, Wang C Y, Bi T, Shi Y, Zhou X D and Chen T. 2019. A novel apex-time network for cross-dataset micro-expression recognition// Proceedings of the 8th international conference on affective computing and intelligent interaction (ACII). Cambridge, UK: IEEE: 1-6 [DOI: 10.1109/ACII.2019.8925525]

See J, Yap M H, Li J T, Hong X P and Wang S J. 2019. MEGC 2019—the second facial micro-expressions grand challenge//Proceedings of the 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019). Lille, France: IEEE: 1-5 [DOI: 10.1109/FG.2019.8756611]

Song B L, Li K, Zong Y, Zhu J, Zheng W M, Shi J G and Zhao L. 2019. Recognizing spontaneous micro-expression using a three-stream convolutional neural network. IEEE Access, 7: 184537-184551 [DOI: 10.1109/ACCESS.2019.2960629]

Van Quang N, Chun J and Tokuyama T. 2019. CapsuleNet for micro-expression recognition//Proceedings of the 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019). Lille, France: IEEE: 1-7 [DOI: 10.1109/FG. 2019. 8756544]

Wei M T, Zheng W M, Zong Y, Jiang X X, Lu C and Liu J T. 2022. A novel micro-expression recognition approach using attention-based magnification-adaptive networks//Proceedings of 2022 ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 2420-2424 [DOI: 10.1109/ICASSP43922.2022.9747232]

Xia Z Q, Peng W, Khor H Q, Feng X Y and Zhao G Y. 2020. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. IEEE Transactions on Image Processing, 29: 8590-8605 [DOI: 10.1109/TIP.2020.3018222]

Xu F, Zhang J P and Wang J Z. 2017. Microexpression identification and categorization using a facial dynamics map. IEEE Transactions on Affective Computing, 8(2): 254-267 [DOI: 10.1109/TAFFC.2016.2518162]

Yan W J, Wu Q, Liang J, Chen Y H and Fu X L. 2013. How fast are the leaked facial expressions: the duration of micro-expressions. Journal of Nonverbal Behavior, 37(4): 217-230 [DOI: 10.1007/s10919-013-0159-8]

Zhai Z J, Zhao J H, Long C J, Xu W J, He S J and Zhao H J. 2023. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22086-22095 [DOI: 10.1109/CVPR52729.2023.02115]

Zhao G Y and Pietikainen M. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6): 915-928 [DOI: 10.1109/TPAMI.2007.1110]

Zhao M H, Dong S S, Hu J, Du S L, Shi C, Li P and Shi Z H. 2024. Attention-guided three-stream convolutional neural network for microexpression recognition. Journal of Image and Graphics, 29(1): 111-122 (赵明华, 董爽爽, 胡静, 都双丽, 石程, 李鹏, 石争浩. 2024. 注意力引导的三流卷积神经网络用于微表情识别. 中国图象图形学报, 29(1): 111-122) [DOI: 10.11834/jig.230053]

Zhao S R, Tang H Y, Liu S F, Zhang Y S, Wang H, Xu T, Chen E H and Guan C T. 2022. ME-PLAN: a deep prototypical learning with local attention network for dynamic micro-expression recognition. Neural Networks, 153: 427-443 [DOI: 10.1016/j. neunet. 2022. 06.024]

Zhou L, Mao Q R, Huang X H, Zhang F F and Zhang Z H. 2022. Feature refinement: an expression-specific feature learning and fusion method for micro-expression recognition. Pattern Recognition, 122: #108275 [DOI: 10.1016/j.patcog.2021.108275]

Zong Y, Huang X H, Zheng W M, Cui Z and Zhao G Y. 2018. Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. IEEE Transactions on Multimedia, 20(11): 3160-3172 [DOI: 10.1109/TMM.2018.2820321]

作者简介

沈天舒,女,硕士研究生,主要研究方向为微表情识别及图像处理。E-mail:sts9842@126.com

迟静,通信作者,女,教授,主要研究方向为计算机动画、图形图像智能处理及大数据建模。E-mail:chijing@sdufe.edu.cn

王雁冰,女,硕士研究生,主要研究方向为人脸图像复原。E-mail:842325005@qq.com

雷雁蕾,男,硕士研究生,主要研究方向为语音驱动人脸动作序列生成。E-mail:leiyanlei@mail.sdufe.edu.cn

徐铭,男,硕士研究生,主要研究方向为三维人脸重建。E-mail:993910364@qq.com