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
©中国图象图形学报版权所有

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)10-3060-14

论文引用格式:Bao Y T, Wu C X, Zhang P and Shan C F. 2024. Adaptive optical flow estimation-driven micro-expression recognition. Journal of Image and Graphics, 29(10):3060-3073(包永堂, 武晨曦, 张鹏, 单彩峰. 2024. 自适应光流估计驱动的微表情识别. 中国图象图形学报, 29(10):3060-3073)[DOI:10.11834/jig.230566]

自适应光流估计驱动的微表情识别

包永堂[1],武晨曦[1],张鹏[1*],单彩峰[2]

1. 山东科技大学计算机科学与工程学院,青岛　266590; 2. 山东科技大学电气与自动化学院,青岛　266590

摘　要:**目的**　微表情识别旨在从面部肌肉应激性运动中自动分析和鉴别研究对象的情感类别,其在谎言检测、心理诊断等方面具有重要应用价值。然而,当前微表情识别方法通常依赖离线光流估计,导致微表情特征表征能力不足。针对该问题,提出了一种基于自适应光流估计的微表情识别模型(adaptive micro-expression recognition, AdaMER)。**方法**　AdaMER并行联立实现光流估计和微表情分类两个任务自适应学习微表情相关的运动特征。首先,提出密集差分编码—解码器以提取多层次面部位移信息,实现自适应光流估计;然后,借助视觉Transformer挖掘重建光流蕴含的微表情判别性信息;最后,融合面部位移微表情语义信息与微表情判别信息进行微表情分类。**结果**　在由SMIC(spontaneous micro-expression recognition)、SAMM(spontaneous micro-facial movement dataset)和CASME II(the Chinese Academy of Sciences micro-expression)构建的复合微表情数据集上进行大量实验,结果显示本文方法UF1(unweighted F1-score)和UAR(unweighted average recall)分别达到了82.89%和85.95%,相比于最新方法FRL-DGT(feature representation learning with adaptive displacement generation and Transformer fusion)分别提升了1.77%和4.85%。**结论**　本文方法融合了自适应光流估计与微表情分类两个任务,一方面以端到端的方式实现自适应光流估计以感知面部细微运动,提高细微表情描述能力;另一方面,充分挖掘微表情判别信息,提升微表情识别性能。

关键词:微表情识别;自适应光流估计;运动特征;差分编码;特征融合

Adaptive optical flow estimation-driven micro-expression recognition

Bao Yongtang[1], Wu Chenxi[1], Zhang Peng[1*], Shan Caifeng[2]

1. College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao　266590, China;

2. College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao　266590, China

Abstract: **Objective** Micro-expressions are brief, subtle facial muscle movements that accidentally signal emotions when the person tries to hide their true inner feelings. Micro-expressions are more responsive to a person's true feelings and motivations than macro-expressions. Micro-expression recognition aims to analyze and identify automatically the emotional category of the research object from the stressful movement of the facial muscles, which has an important application value in lie detection, psychological diagnosis, and other aspects. In the early development of micro-expression recognition, local binary patterns and optical flow were widely used as features for training traditional machine learning models. However, the traditional manual feature approach relies on manually designing rules, making it difficult to adapt to the differences in

收稿日期:2023-08-18;**修回日期**:2023-12-21;**预印本日期**:2023-12-28

*通信作者:张鹏　pengzhang_skd@sdust.edu.cn

基金项目:山东省自然科学基金项目(ZR2021QF017, ZR2020MF132);山东科技大学青年教师教学拔尖人才培养项目(2023BJ1201)

Supported by: Natural Science Foundation of Shandong Province, China(ZR2021QF017, ZR2020MF132)

micro-expression data across different individuals and scenarios. Given that deep learning can automatically learn the optimal feature representation of an image, the recognition performance of micro-expression recognition studies based on deep learning far exceeds that of traditional methods. However, micro-expressions occur as subtle facial changes, which causes the micro-expression recognition task to remain challenging. By analyzing the pixel movement between consecutive frames, the optical flow can represent the dynamic information of micro-expressions. Deep learning-based micro-expression recognition methods perform facial muscle motion descriptions with optical flow information to improve micro-expression recognition performance. However, existing micro-expression recognition methods usually extract the optical flow information offline, which relies on existing optical flow estimation techniques and suffers from the insufficient description of subtle expressions and neglect of static facial expression information, which restricts the recognition effect of the model. Therefore, this study proposes a micro-expression recognition network based on adaptive optical flow estimation, which realizes optical flow estimation and micro-expression classification to learn micro-expression-related motion features through parallel association adaptively. **Method** The training samples of micro-expressions are limited, which makes it difficult to train complex network models. Therefore, this study selects the apex and their neighboring frames in the micro-expression video sequence as training data in the preprocessing stage. In addition, when loading the data, the original training data are replaced with image pairs containing motion information in the video sequence with a certain probability. Second, the deep learning network with a dense differential encoder-decoder implements the facial muscle motion adaptive optical flow estimation task to improve the characterization of subtle expressions. ResNet18 extracts features from the two-frame image and the difference map in a dense differential encoder. The branch processing the two frames shares the parameters. A motion enhancement module is added to the feature extraction branch of the differential image to accomplish the interlayer information interaction. In the motion enhancement module, the difference map features computed from the two frames need the spatial attention mechanism to focus on the micro-expression-related motion; the two frames are subtracted from each other to preserve and amplify the difference between the two frames, and using the two features provides valid information for subsequent networks. The decoder in this study maps the multilevel facial displacement information extracted by the dense differential encoder and the last layer of the two-frame image output features to reconstruct the optical flow features. Vision Transformer is a deep learning model based on the self-attention mechanism, which has global perception capability in comparison with the traditional convolutional neural network. Then, with the feature extraction capability of vision Transformer, the micro-expression discriminative information embedded in the reconstructed optical flow is mined. Finally, the semantic information of micro-expressions extracted from facial displacement information and the discriminative information of micro-expressions extracted from the vision Transformer model are fused to provide rich information for micro-expression classification. This study uses the Endpoint error loss constraint for the optical flow estimation task to achieve the learning purpose, which continuously reduces the Euclidean distance between the predicted and real optical flow. Cross entropy loss function constraints are used for the features extracted by vision Transformer and the fused features, which make the network learn micro-expression related information. At the same time, the image with low motion intensity in the two frames is equivalent to the neutral expression (without motion information), and the KL-divergence loss is applied to the output of the feature by the encoder to suppress irrelevant information. The loss functions interact to complete the network optimization. **Result** This study evaluates the model performance on a public dataset using the leave-one-subject-out cross-validation evaluation strategy. Face alignment and cropping are performed on the public dataset samples to unify the dataset. To demonstrate the state-of-the-art of the proposed method, we compare it with existing mainstream methods on composite datasets constructed by SMIC, SAMM, and CASME II. Our method achieves 82.89% and 85.59% UF1 and UAR on the whole dataset, 78.16% and 80.89% UF1 and UAR on the SMIC part, 94.52% and 96.02% UF1 and UAR on the CASME II part, and 73.24% and 75.83%. Our method achieves optimal results in the whole dataset, the SMIC part, and the CASME II part, and suboptimal results in the SAMM part. Compared to the latest proposed micro-expression method based on feature representation learning with adaptive displacement generation and Transformer fusion (FRL-DGT), our method demonstrates an improvement of 1.77% and 4.85%. **Conclusion** The micro-expression recognition model based on adaptive optical flow estimation proposed in this study fuses the proposed two tasks of adaptive optical flow estimation and micro-expression categorization, which, on the one hand, senses the subtle facial movements in an end-to-end manner and

improves the ability of subtle expression description, and on the other hand, fully exploits the micro-expression discriminative information and enhances the micro-expression performance.

Key words: micro-expression recognition; adaptive optical flow estimation; motion features; differential encoder; feature fusion

0 引 言

微表情是人内心真实情感的外在映射，以面部肌肉应激形式呈现，表现为当一个人试图压抑或者管理真实情感时意外流露的情绪信号，通常仅持续1/25～1/3 s（Ekman和Friesen，1975）。与宏观表情相比，微表情难以掩盖和控制，更能传达人真实的内心情感（Bhushan，2015；陈港 等，2022），因此广泛应用于谎言检测和心理学诊断等领域。然而，微表情持续时间短且表达强度低，须经过专门培训的心理学专家才能鉴别和区分，这极大限制了其应用范围（Ben等，2022；Sun等，2022）。随着人工智能领域的高速发展，近年来研究人员尝试利用计算机视觉、多媒体等智能技术自动定位和识别微表情。

目前，微表情识别方法可以归纳为两类：基于特征工程的方法和基于数据驱动学习的方法。前者旨在设计特征描述微表情发生时面部纹理、运动等信息，例如局部二值模式（Zhao和Pietikainen，2007）、光流（Shreve等，2011）等，是早期微表情识别研究的热点。然而，传统微表情特征设计严重依赖于经验，难以充分挖掘微表情特征，性能无法满足实际微表情识别系统的需求。随着深度学习在计算机视觉领域的巨大成功，基于数据驱动学习的微表情识别引起了许多研究人员的关注。部分研究（Li等，2018，2021；Van Quang等，2019；Yu等，2021；Lei等，2020，2021）使用神经网络模型直接从视觉图像或视频中面部强度最大的顶点帧学习微表情相关语义特征。然而，微表情强度非常细微，单纯依赖顶点帧通常难以获取足够的判别特征进行微表情分类。为了克服该问题，大量工作（Khor等，2019；Li等，2021；Su等，2021；Zhou等，2022；Zhang等，2022；Chen等，2023；Alkaddour等，2022）关注微表情发生时面部肌肉运动引起的局部区域动态变化，提出基于光流信息的微表情识别方法。这类方法一定程度上可以减少静态图像中身份信息对微表情识别的干扰，取得了目前最好的性能，证明了光流在微表情识别中的有效性。然而，这些工作依赖于现有通用光流估计方法预先提取光流，无法准确且充分描述细微表情变化，限制了该类方法的性能。

针对已有光流估计方法的不足，微表情识别研究最新工作（Nguyen等，2023；Fan等，2023；Zhai等，2023）展开微表情自适应动态特征学习，通过分析帧间像素位移学习微表情发生时面部肌肉运动特征，提高了微表情的识别性能。然而，上述方法通常以自监督的方式约束微表情动态信息生成，缺少外部知识的引入和监督，受限于微表情数据增广方式。

与上述方法不同，本文提出了一种外部知识监督的融合自适应光流估计的微表情识别模型（adaptive micro-expression recognition，AdaMER）。该模型将现有光流方法作为外部监督知识约束动态特征的生成，同时以微表情识别模型作为动态特征校正和微表情信息注入网络，约束生成的动态特征蕴含微表情信息，通过端到端的方式同时实现面部运动光流估计和微表情识别。特别地，为了增强面部微小局部运动的感知能力，本文提出了一种融合运动增强模块（motion enhancement module，MEM）的密集差分编码器，采用跨尺度运动（差分信息）传递机制结合不同尺度帧间差分信息，逐渐增强面部细微运动蕴含的微表情语义信息，克服现有光流估计方法（Dosovitskiy 等，2015；Ilg 等，2017；Teed 和 Deng，2020；Jiang等，2021）对细微运动不敏感，防止细微运动信息在深度网络传输过程中发生信息丢失的问题。此外，受Chen等人（2023）启发，本文提出方法对微表情帧间动态特征和帧内静态视觉特征进行融合，提高特征表达能力。

本文主要贡献如下：1）针对面部肌肉运动特点，提出了一种密集差分编码器，通过运动增强模块实现不同尺度面部差分信息的交互和聚合，学习面部肌肉运动。2）提出端到端的面部运动光流估计与微表情识别多任务学习统一模型，以光流估计作为正则化条件，通过引入外部光流知识指导信息，同时实现自适应光流估计与微表情分类，增强微表情分类模型的泛化能力。3）为了增强模型的鲁棒性，将卷

积块组成的编码器提取微表情语义特征和视觉Transformer(Dosovitskiy等,2021)提取判别性特征进行特征融合,提高特征表达能力,进一步提高微表情识别性能。

本文在3个公开数据集SMIC(spontaneous micro-expression database)、CASME II(the Chinese Academy of Sciences micro-expression)和SAMM(spontaneous micro-facial movement dataset)上进行了大量的实验,结果证实了本文方法的有效性。与基于传统光流估计方法的微表情识别模型相比,所提出的网络结构在微表情识别上实现了更优的识别性能。

1 相关工作

根据微表情特征提取方式的不同,现有微表情识别工作可归纳为两类:基于传统手工特征工程的方法和基于数据驱动学习的方法。

1.1 基于传统手工特征工程的方法

早期的微表情识别工作主要聚焦于微表情特征描述符人工设计,严重依赖研究者的经验。例如,Zhao和Pietikainen(2007)提出使用局部二值模式(local binary patterns,LBP)描述面部纹理细节,结合传统机器学习分类器实现微表情自动识别。然而,LBP只能表征图像信息,对微表情视频序列表示能力不足。为了解决上述问题,研究人员提出了一系列LBP改进算法,包括Yan等人(2014)提出的LBP-TOP(local binary patterns from three orthogonal planes)和Wang等人(2014)提出的LBP-SIP(local binary patterns with six intersection points)等。然而,这类方法通常只能描述静态图像信息,无法挖掘微表情动态特征。为了解决这个问题,基于光流的微表情动态特征描述方法被提出,例如,Shreve等人(2011)使用中心差分法计算微表情发生时建立在下颌、嘴巴、脸颊和额头4个区域密度光流场的肌肉张力幅度;Liu等人(2016)提出使用MDMO(main directional mean optical-flow)对运动信息和空间位置进行局部计算;See等人(2019)提出Bi-WOOF(bi-weighted oriented optical flow),计算起始帧和顶点帧的光流图并使用光学应变和光流幅度进行加权获得光流直方图特征。然而,这些手工特征在很大程度上依赖于研究人员的经验,无法完全挖掘微表情特征,性能难以满足实际微表情识别系统的需求。

1.2 基于数据驱动学习方式的方法

随着深度学习在计算机视觉领域的巨大成功,基于数据驱动学习的微表情识别近年来引起了研究人员的大量关注。这些工作通常可以分为两类:一类是基于视觉特征提取的方法;另一类是基于动态特征学习的方法。

基于视觉特征的方法直接从视觉图像或视频中学习微表情相关特征,例如,Li等人(2018)提出利用频域信息检测微表情序列顶点帧,并使用卷积神经网络提取顶点帧内微表情信息,证实了仅使用运动强度最大的顶点帧就可以达到良好的性能。基于此,Li等人(2021)构建了顶点帧框架,同时利用局部和全局特征,表现出良好的性能。为了获得区分性特征,Lei等人(2020,2021)对微表情顶点帧涉及的面部运动区域之间的关系进行建模,通过图表示学习提取微表情信息。然而,微表情强度非常细微,单纯依赖顶点帧通常难以获取足够的判别特征进行微表情分类。

为了充分利用面部运动特征,基于动态特征学习的方法被提出,旨在学习微表情发生时面部肌肉变化的动态特征。例如,Khor等人(2019)提取微表情起始帧和顶点帧之间的光流,通过双流浅层神经网络(dual-stream shallow networks, DSSN)学习光流幅值和光流应变中的微表情特征信息。类似地,Su等人(2021)提出计算起始帧和顶点帧之间的一阶光流和二阶光流,结合组件感知注意力提取微表情感兴趣区域判别特征。Zhou等人(2022)提出双流Inception网络(dual inception)学习光流水平和垂直方向互补特征,提高微表情表征的判别能力。为了捕获局部和全局时空特征,Zhang等人(2022)提出基于视频序列的长短依赖的时空Transformer网络(short and long range relation based spatio-temporal Transformer,SLSTT)。为了缓解微表情数据集小样本问题,Chen等人(2023)提出基于已有光流提取方法生成4种类型的光流增强微表情样本。这些方法一定程度上可以减少静态图像中身份信息对微表情识别的干扰,取得了目前最好的性能,证明了光流在微表情识别中的有效性。然而,这些工作依赖于现有通用光流估计方法预先提取光流,并不适用于提取微表情发生时的细微表情变化信息。

考虑到基于已有光流估计技术的微表情识别工作的局限性,研究人员开始将自适应学习的思想用

于微表情识别工作，Nguyen等人(2023)提出一种对角线微注意力(diagonal micro-attention，DMA)机制，通过学习两帧图像间局部区域互换的细微差异模拟微表情动态特征学习。尽管这种方法一定程度上可以学习微表情动态特征，其局部区域互换受限于两帧图像间的对齐。与之不同，Fan等人(2023)提出一种基于关键点检测构建面部位移图像的方法，同时通过对称对比视觉Transformer(symmetric contrastive vision Transformer，SCViT)以自监督方式学习面部左右两侧对称的面部动作。然而，该方法依赖于准确的关键点定位，且忽略了微表情静态视觉特征信息。Zhai等人(2023)提出一种融合自适应位移生成与Transformer特征学习的深度学习网络(feature representation learning with adaptive displacement generation and Transformer fusion，FRL-DGT)，通过生成模型学习起始帧与顶点帧间位移信息，并结合两帧内视觉静态信息进行微表情识别。借助生成模型的推理能力，这些方法能够自适应学习面部动态特征，提高了微表情识别性能。然而，上述方法通常以自监督方式约束微表情动态信息生成，缺少外部知识的引入和监督，受限于微表情视角增广方式。受到郑壮强等人(2020)启发，本文提出一种外部知识监督的融合自适应光流估计的微表情识别模型，旨在通过端到端的方式同时实现面部运动光流估计和微表情识别。

2 本文方法

如图1所示，本文提出基于自适应光流估计的微表情识别模型AdaMER，由数据预处理、微表情自适应光流估计与学习和表情分类3部分组成。数据预处理部分包括帧采样和基准光流生成。具体地，给定一段微表情视频序列，帧采样指从视频序列选择起始帧与顶点相邻5帧间构成的图像对，表示为($\boldsymbol{X}_1$, $\boldsymbol{X}_2$)。对于采样的两帧，本论文采用TV_L1(Zach等，2007)计算帧间光流作为外部监督知识，引导模型生成光流信息。微表情光流估计部分由一个密集差分编码器—解码器构成，通过计算不同层内顶点帧和起始帧间的差分信息，学习微表情动态位移信息，实现微表情光流估计和重构。微表情识

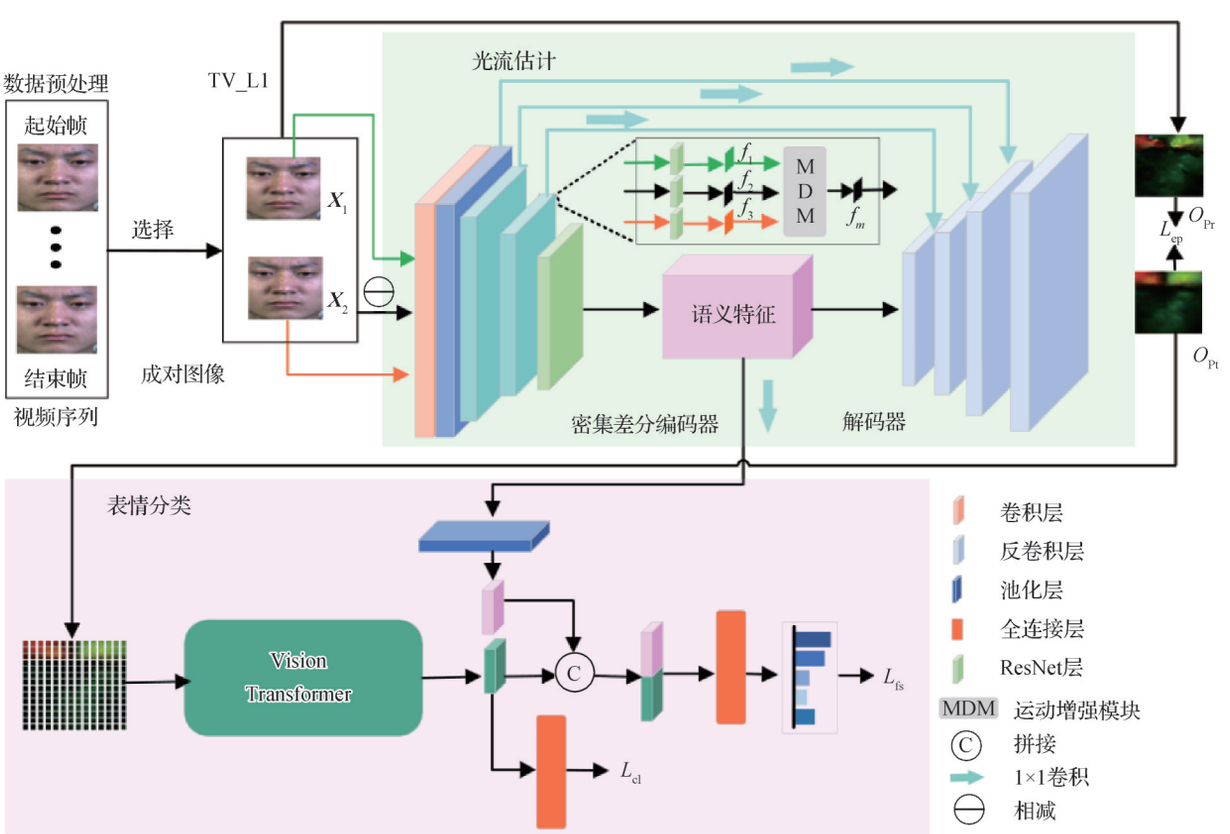


图1 自适应微表情识别模型整体框架

Fig. 1 Overall framework of the adaptive micro-expression recognition (AdaMER) model

别部分包含一个视觉Transformer特征提取网络和微表情分类器。视觉Transformer学习预测光流包含的运动信息，同时与差分编码器输入动态以及静态特征融合，通过微表情分类器进行微表情类别预测。

2.1 自适应微表情光流估计与学习

为了提取微表情发生时面部运动特征，本文提出了一种基于密集差分编码器—解码器的面部肌肉细微运动光流估计模型，自适应学习微表情发生时帧间动态变化信息。

2.1.1 密集差分编码器

根据光流估计原理，本文提出的密集差分编码器（dense differential encoder，DSDE）自适应学习微表情发生时帧间水平和垂直方向位移信息，具体如图2所示。根据已有的微表情研究（Khor等，2019；李国豪 等，2020；Su等，2021；Zhou等，2022；Zhang等，2022；Chen等，2023），给定微表情发生时顺序采样的两帧图像$\boldsymbol{X}_1$和$\boldsymbol{X}_2$，二者差分信息反映了微表情发生时面部肌肉运动导致的像素变化，蕴含了微表情的运动特征，即

$$\boldsymbol{X}_3 = BN(\boldsymbol{X}_2 - \boldsymbol{X}_1) \tag{1}$$

式中，$\boldsymbol{X}_3$表示差分图像，BN表示批量归一化。本文提出密集差分编码器包含三路分支网络，分别提取起始帧、顶点帧和差分帧的表情特征，即

$$\boldsymbol{f}_1, \boldsymbol{f}_2 = F_l(\boldsymbol{X}_1, \boldsymbol{X}_2) \tag{2}$$

$$\boldsymbol{f}_3 = G_l(\boldsymbol{X}_3) \tag{3}$$

式中，$F_l(\cdot)$和$G_l(\cdot)$分别表示共享分支与差分分支第l个卷积层操作。值得注意的是，本文采用ResNet18（residual network 18 layers）（He等，2016）前4层作为密集差分编码器三路分支的骨干网络，$\boldsymbol{f}_1$，$\boldsymbol{f}_2$分别是$\boldsymbol{X}_1$、$\boldsymbol{X}_2$经过共享参数分支后输出的特征，用于提取面部静态特征；$\boldsymbol{f}_3$是$\boldsymbol{X}_3$经过差分分支后输出的特征，用于提取面部动态特征。

2.1.2 运动增强模块

为了防止微表情差分信息在网络传播过程中丢失，本文设计了一个运动增强模块（motion enhancement module，MEM），通过层内静态特征密集差分增强差分偏移特征，其输出表示为

$$\boldsymbol{f}_m = M(\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3) \tag{4}$$

式中，$M(\cdot)$代表运动增强模块的学习过程，如图3所示，运动增强模块将起始帧静态特征$\boldsymbol{f}_1$、顶点帧静态特征$\boldsymbol{f}_2$以及动态特征$\boldsymbol{f}_3$融合，增强微表情差分信息。具体地，MEM首先计算每层内起始帧和顶点帧间特征差分信息，以捕捉不同尺度视觉特征的差异增强微表情差分特征，即

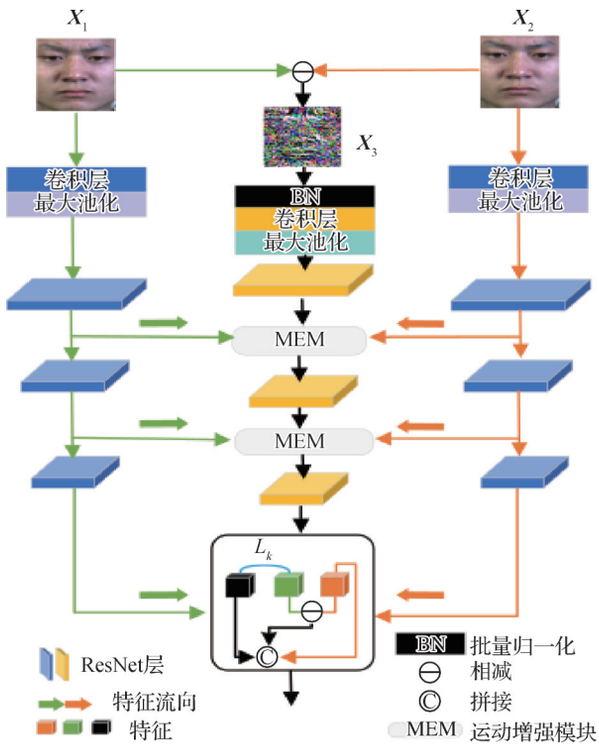


图2 密集差分编码器结构

Fig. 2 The structure dense differential encoder structure

$$\boldsymbol{f}_d = BN(\boldsymbol{f}_2 - \boldsymbol{f}_1) \tag{5}$$

式中，$\boldsymbol{f}_d$表示当前层内顶点帧和起始帧间位移动态特征。

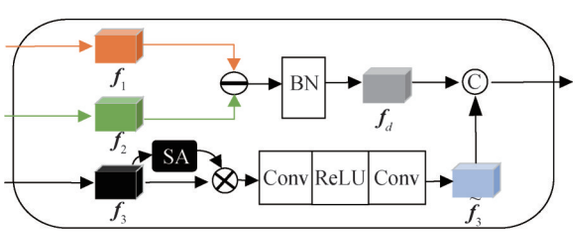


图3 运动增强模块结构

Fig. 3 The structure of motion enhancement module

考虑到差分图像$\boldsymbol{X}_3$中可能包含头部运动等非表情信息，本文采用一个空间注意力模块（spatial attention module，SA）（Woo等，2018）使模型关注微表情发生时面部运动区域。为增强网络的表征能力，模块中使用了卷积核大小为1×1卷积Conv和激活函数ReLU（rectified linear unit），具体表示为

$$\tilde{\boldsymbol{f}}_3 = \mathrm{Conv}(\mathrm{ReLU}(\mathrm{Conv}(\boldsymbol{f}_3' \times \boldsymbol{f}_3)) \tag{6}$$

式中，$\boldsymbol{f}_3'$为SA模块的输出结果，$\tilde{\boldsymbol{f}}_3$为经过特征增强

操作的输出特征。最后，通过拼接操作融合f_d与f_3进行差分信息的增强，即

$$f_m = Concat(f_d, \tilde{f}_3) \tag{7}$$

式中，$Concat(\cdot)$为特征拼接操作，f_m是差分信息增强特征。为了保留帧内静态微表情特征，本文将密集差分编码器最后一层的输出差分位移动态特征f_d、顶点帧静态视觉特征f_2和差分信息增强特征f_m进行拼接融合，得到编码器输出微表情特征f_{md}，表示为

$$f_{md} = Concat(f_2, f_m, f_d) \tag{8}$$

此外，为了充分学习微表情静态特征和动态特征，本文最大化KL（Kullback-Leibler）散度损失来增加最后一层输出静态视觉特征和动态差分特征的差异性，结合微表情光流重构约束和微表情分类约束，使模型尽可能充分挖掘动态信息和静态信息。具体地，通过起始帧视觉特征f_1和差分动态特征f_d计算K1散度损失，即

$$L_k(f_1, f_d) = \sum_{i=1}^{N} f_1 \log\left(\frac{f_1}{f_d}\right) \tag{9}$$

式中，N为小批量中样本数。

2.1.3　解码器

为了使编码器能够提取微表情动态信息，本文提出使用光流重构任务对编码器进行正则化，通过添加一个解码器，使差分编码器提取动态特征能够解码为面部基准光流。为了避免编码过程中信息丢失造成解码不完全问题，采用UNet（Ronneberger等，2015）中的Skip-connection结构，即差分编码器模块输出的多层次位移特征图f_m经过1×1的卷积通道数减半，与上采样的输出的预测结果沿通道维度拼接，然后经过1×1的卷积降维作为下层转置卷积的输出。通过这种特征融合方式保留了低层特征，克服解码过程中信息不足问题。为了使网络能够学习细微变化，本文采用端点误差（endpoint error，EPE）损失计算预测光流和光流真值的欧氏距离，具体为

$$L_{ep} = \sqrt{(\boldsymbol{u}_{pr} - \boldsymbol{u}_{gt})^2 + (\boldsymbol{v}_{pr} - \boldsymbol{v}_{gt})^2} \tag{10}$$

式中，$\boldsymbol{u}_{pr}$和$\boldsymbol{v}_{pr}$分别表示预测光流水平方向的位移和垂直方向位移，$\boldsymbol{u}_{gt}$和$\boldsymbol{v}_{gt}$分别表示基准光流水平方向位移和垂直方向位移。然而，微表情数据集里没有提供基准面部运动光流。为了解决这个问题，引入外部知识作为指导信息，即采用光流估计方法TV_L1提取光流作为基准光流引导模型产生运动信息。虽然TV_L1不能提供精确的基准光流，它引导模型学习面部变化特征，与微表情分类约束共同作用下，引导模型学习微表情相关运动特征。

2.2　微表情特征融合和分类

2.2.1　微表情视觉Transformer模型

面部运动编码系统（facial action coding system，FACS）分析了表情和面部运动单元（action unit，AU）的对应关系（Szajnberg，2022）。根据公开的微表情数据，发现不同的表情对应于不同的面部运动单元，因此部分方法利用运动单元定位提取局部信息。然而，运动单元定位通常依赖面部关键点定位，计算比较复杂。为了解决这个问题，本文采用视觉Transformer模型，借助块信息之间全局依赖关系逼近微表情发生时运动单元。

根据视觉Transformer原理，首先将预测光流图$\boldsymbol{O}_{pr}$分割成不重叠的局部区域，并使用可训练的线性投影将矢量化的局部区域$\boldsymbol{o}_p$映射到潜在的D维嵌入空间，即$\boldsymbol{o}_p\boldsymbol{E}$。为了保留位置信息，本文在局部区域嵌入中引入可学习的位置嵌入，表示为

$$\boldsymbol{O}_0 = [\boldsymbol{o}_c, \boldsymbol{o}_p^1\boldsymbol{E}, \boldsymbol{o}_p^2\boldsymbol{E}, \cdots, \boldsymbol{o}_p^N\boldsymbol{E}] + \boldsymbol{E}_{ps} \tag{11}$$

式中，$\boldsymbol{O}_0$为视觉Transformer的输入，N为光流图像分割后局部区域的个数，$\boldsymbol{E} \in \mathbf{R}^{(P^2 \cdot C) * D}$是局部区域嵌入投影，$C$为图像通道数，$\boldsymbol{E}_{ps} \in \mathbf{R}^{N*D}$是位置嵌入，$\boldsymbol{o}_p^i$是光流图$\boldsymbol{O}_{pr}$分割的第$i$个局部区域矢量化表示，$\boldsymbol{o}_c$是类别Token，用于汇集局部区域特征。由于局部区域在不同的微表情中具有不同的运动强度，使用随机初始化的类别Token可以避免对某一区域过于关注。Transformer层包括l层的多头自注意力和多层感知机组成的块。第l层的输出可表示为

$$\boldsymbol{O}_l' = MSA(LN(\boldsymbol{O}_{l-1})) + \boldsymbol{O}_{l-1}, l \in 1, 2, \cdots, L \tag{12}$$

$$\boldsymbol{O}_l = MLP(LN(\boldsymbol{O}_l')) + \boldsymbol{O}_l', l \in 1, 2, \cdots, L \tag{13}$$

式中，$LN(\cdot)$是层归一化，$\boldsymbol{O}_{l-1}$是第$l-1$层输出的特征，$\boldsymbol{O}_l$是第l层输出的特征，$MSA(\cdot)$表示多头自注意力，$\boldsymbol{O}_l'$是残差连接操作后的输出特征，$MLP(\cdot)$表示多层感知机，L是视觉Transformer总的Transformer层数。由于$\boldsymbol{o}_c$通过自动加权汇集了局部特征，所以将最后一个Transformer层输出$\boldsymbol{o}_c$作为微表情类别特征，即$f_{lt} = \boldsymbol{o}_c$。

2.2.2　特征融合

密集差分编码器输出的语义特征包含来自两帧

图像 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 的静态和动态表情信息。考虑到微表情识别模型性能的提升需要更可靠的微表情相关特征，将视觉 Transformer 提取的特征 $\boldsymbol{f}_{\mathrm{lt}}$ 和编码器输出的语义特征进行融合，具体为

$$\boldsymbol{f}'_{\mathrm{md}} = Flatten(Maxpool(Conv(\boldsymbol{f}_{\mathrm{md}}))) \tag{14}$$

$$\boldsymbol{f}_{\mathrm{fs}} = Concat(\boldsymbol{f}'_{\mathrm{md}}, \boldsymbol{f}_{\mathrm{lt}}) \tag{15}$$

式中，$Flatten(\cdot)$ 是张量拉平操作，$\boldsymbol{f}'_{\mathrm{md}}$ 是密集差分编码器输出的语义特征经转化后的输出，$\boldsymbol{f}_{\mathrm{fs}}$ 是用来表情预测的特征。

2.3 优化目标

对于微表情分类，本文在视觉 Transformer 输出的全局特征 $\boldsymbol{f}_{\mathrm{lt}}$ 位置添加了全连接层构成的分类器，并用交叉熵损失引导模型产生微表情判别特征，具体为

$$L_{\mathrm{cl}} = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|FC(\boldsymbol{f}_{\mathrm{lt}})) \tag{16}$$

式中，$FC(\cdot)$ 表示全连接层操作，$p(y_i|FC(\boldsymbol{f}_{\mathrm{lt}}))$ 是特征 $FC(\boldsymbol{f}_{\mathrm{lt}})$ 预测为对应真实类别 y_i 的概率。同时，对融合特征 $\boldsymbol{f}_{\mathrm{fs}}$ 添加全连接层构成的分类器，获得微表情识别模型的分类结果。具体为

$$L_{\mathrm{fs}} = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|FC(\boldsymbol{f}_{\mathrm{fs}})) \tag{17}$$

式中，$p(y_i|FC(\boldsymbol{f}_{\mathrm{fs}}))$ 是特征 $FC(\boldsymbol{f}_{\mathrm{fs}})$ 预测为对应真实类别 y_i 的概率。为了同时实现微表情光流估计与微表情识别，本文同时训练微表情重构网络和微表情分类网络。因此，本文提出的 AdaMER 的总体损失为

$$L_t = L_{\mathrm{ep}} + \lambda_1 L_{\mathrm{cl}} + \lambda_2 L_{\mathrm{fs}} + \lambda_3 L_k \tag{18}$$

式中，λ_1、λ_2 和 λ_3 是平衡相应部分贡献的超参数。通过最小化总损失 L_t，使 AdaMER 同时具备微表情动态光流预测和微表情判别特征提取的能力。

3 实验结果与分析

3.1 实验数据

为了与现有的工作进行公平比较，本文在 SMIC（Li 等，2013）、CASME II（Yan 等，2014）和 SAMM（Davison 等，2018）3 个基准数据库上进行实验。与现有方法相同，采用留一被试（leave-one-subject-out，LOSO）交叉验证策略，每个受试者都被用做测试集，其余样本用做训练集。

SMIC 数据集提供了 3 种不同类型的录制摄像机：高速摄像机、通用视觉摄像机和近红外相机。由于高速摄像机可以捕捉微表情的瞬时变化，微表情实验中选用高速摄像机录制的样本。数据集包含的样本来自 16 个参与者，分辨率为 640 × 480 像素，包含 3 种情绪：积极（positive）、消极（negative）和惊讶（surprise）。

CASME II 包含相机以 200 帧/s 的速度记录的 247 个样本。26 名参与者在一个控制良好的实验环境中生成了分辨率为 640 × 480 像素的样本。此外，CASME II 包含 5 种情绪：高兴（happiness）、压抑（repression）、惊讶（surprise）、厌恶（disgust）和其他（other）。所有样本包含起始帧、顶点帧、偏移帧和面部运动单元注释。

SAMM 包含 159 个由高速相机以 200 帧/s 的速度记录的样本，分辨率为 2 040 × 1 088 像素。SAMM 有 7 种情绪：高兴（happiness）、惊讶（surprise）、生气（anger）、厌恶（disgust）、悲伤（sadness）、害怕（fear）和其他（others）。每个参与者都需要完成一份问卷和自我报告，以准确地标记表情。与 CASME II 一样，所有的样本都包含起始帧、顶点帧和面部运动单元注释。

3.2 实验设置和评估标准

3.2.1 实验设置

为获得稳健的微表情识别模型，本文对实验数据进行人脸图像预处理。除了 SAMM，SMIC 和 CASME II 数据集提供了裁剪图像，SAMM 进行相应的人脸对齐和裁剪操作，实现数据统一。由于 SMIC 数据集不标注顶点帧，本文使用 CapsuleNet（Van Quang 等，2019）提到的顶点帧标注结果。此外，为了防止模型过拟合，在训练数据加载过程以一定概率使用微表情视频序列随机选择的包含运动信息的两帧替换起始帧和顶点帧图像对。

本文实验环境使用 CPU 为 Intel®Xeon（R）4210R CPU @ 2.4 GHz × 20，操作系统为 Ubuntu 20.04，GPU 为 NVIDIA GeForce RTX 3090。训练过程中，复合数据集和 CASME II 数据集上的批量大小分别设置为 30 和 40。所有实验均采用 Adam 优化器优化网络参数，初始学习率设置为 10^{-5}，随机梯度衰减动量设置为 0.9，总共迭代 100 次。在实验中，通过以下方式实现 AdaMER。首先，将所有用于训练的图像对大小调整为 224 × 224 像素，然后根据光流

估计算法TV_L1(Zach等,2007)计算光流真值。为了保证模型具有提取运动特征的能力,在复合数据集上进行光流估计预训练,然后再同时进行光流估计和表情分类任务提取微表情相关的运动特征。整个框架选择使用参数量较少的ResNet-18的前四层构成编码器,以及深度为2和头部数量为8的视觉Transformer提取预测光流的特征,损失函数的超参数λ_1、λ_2和λ_3分别设置为0.7、0.7和0.0002。

3.2.2 实验评估指标

为了证明提出模型的有效性,本文与已有的微表情识别研究采用相同的评估标准。针对单数据集评估中使用了两个性能指标准确率ACC(accuracy)和宏F1(macroF1)。在复合数据库评估实验中,根据MEGC2019(micro-expressions grand challenge)(See等,2019)提出的标准,通过将CASME II和SAMM数据集的表情重新标记为与SMIC相同的3个常见情绪标签,构建了由3个数据集样本组成的复合数据集,并采用非加权F1分数(unweighted F1-score,UF1)和非加权召回率(unweighted average recall,UAR)两个性能度量指标评估模型的有效性。

3.3 性能比较实验

为了证明所提算法的有效性,本节在复合数据集上与当前最先进的方法进行了实验比较,包括基于传统光流的方法Bi-WOOF(See等,2019)、集合传统光流估计与深度特征学习的方法OFF-ApexNet(apex frame network)(Gan等,2019)、Dual Inception(Zhou等,2019)、RCN(recurrent neural network)(Xia等,2020)、FeatRef(feature refinement)(Zhou等,2022)、SLSTT(short and. Long range relation based spatio-temporal Transformer)(Zhang等,2022)和MSMMT(multi-scale multi-modality Transformer network)(Wang等,2023)以及基于自适应位移生成的微表情识别模型FRL-DGT(Zhai等,2023)。

由实验结果表1可知,本文提出的AdaMER在复合数据集上性能最优,尤其是在CASME II上UF1和UAR指标远超于现有方法。具体来讲,本文提出的模型与传统方法Bi-WOOF相比,整个复合数据集的指标UF1和UAR分别提升了19.93%和23.68%,证明了深度学习在分类任务上具有显著优势。与OFF-ApexNet、Dual Inception和FeatRef基于起始帧和顶点帧之间光流水平方向位移和垂直方向位移的双流模型相比,本文模型提出基于起始帧和顶点帧进行自适应光流估计学习帧间运动特征的方法更适用于微表情识别。FeatReaf不同于Dual Inception,使用注意力机制让模型关注微表情有效的判别性特征。本文模型将注意力机制用于浅层网络ResNet18进行特征选择,与FeatReaf相比,UF1提升4.51%,UAR提升7.63%,证明了本文所提特征提取方式的有效性。SLSTT是第1个在微表情识别上使用纯Transformer的研究,本文方法UF1指标比其高1.29%,UAR指标高6.95%,这证实了融合CNN和

表1 不同网络在复合数据集上对比实验结果

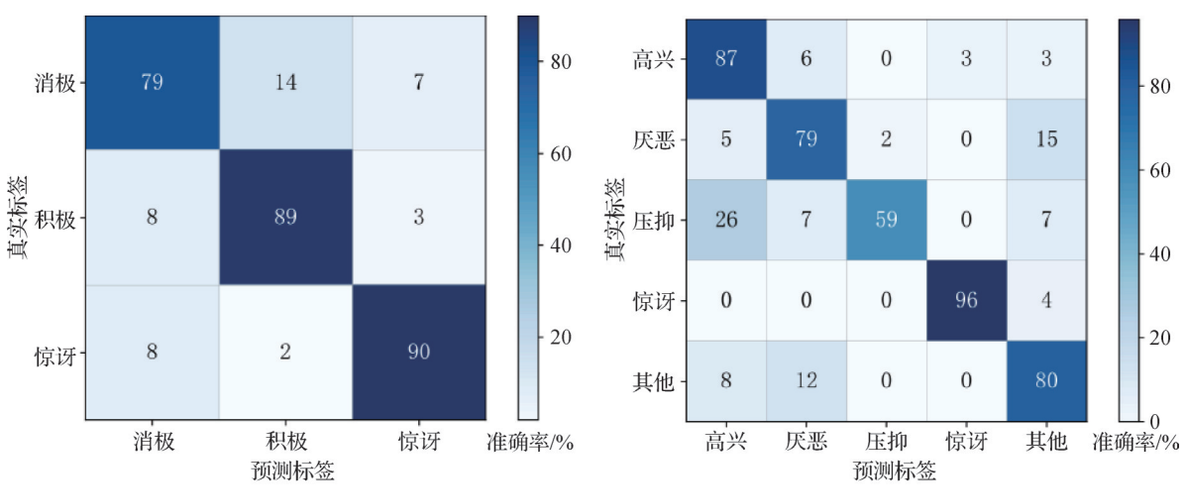
Table 1 Experimental results comparing different networks on composite datasets

网络模型	FULL		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
Bi-WOOF(See等,2019)	0.629 6	0.622 7	0.572 7	0.582 9	0.780 5	0.802 6	0.521 1	0.513 9
OFF-ApexNet(Gan等,2019)	0.719 6	0.709 6	0.681 7	0.695 0	0.876 4	0.868 1	0.540 9	0.530 0
Dual Inception(Zhou等,2019)	0.732 2	0.727 8	0.664 5	0.672 6	0.862 1	0.856 2	0.586 8	0.566 3
RCN(Xia等,2020)	0.705 2	0.716 4	0.598 0	0.599 1	0.808 7	0.856 3	0.677 1	0.697 6
FeatRef(Zhou等,2022)	0.783 8	0.783 2	0.701 1	0.708 3	0.891 5	0.887 3	0.737 2	0.715 5
SLSTT(Zhang等,2022)	0.816 0	0.790 0	0.724 0	0.707 0	0.901 0	0.885 0	0.715 0	0.642 0
MSMMT(Wang等,2023)	0.816 0	0.819 1	0.765 1	0.778 0	0.907 1	0.887 8	0.739 2	0.716 3
FRL-DGT(Zhai等,2023)	0.812 0	0.811 0	0.743 0	0.749 0	0.919 0	0.903 0	**0.772 0**	0.758 0
AdaMER(本文)	**0.828 9**	**0.859 5**	**0.781 6**	**0.808 9**	**0.945 2**	**0.960 2**	0.732 4	**0.758 3**

注:加粗字体表示各列最优结果,FULL为整个复合数据集。

Transformer 提取的特征能提供更丰富的微表情相关特征。与采用多尺度 Transformer 提取动态图像纹理特征和光流图像运动特征的 MSMMT 相比，UF1 指标提升了 1.29%，UAR 指标提升了 4.04%，说明了本文利用密集差分器提供的面部位移语义特征和视觉 Transformer 提取位置等浅层特征帮助区分微表情类别。与自适应学习位移的 FRL-DGT 模型相比，本文方法在未使用运动放大技术的情况下，UF1 提升了 1.77%，UAR 提升了 4.85%，说明了本文提出的密集差分编码器能提取到细微的运动特征。另外，本文方法与传统的光流估计技术 TV_L1 进行光流生成运行时间的比较，从数据集中读取两幅包含运动信息的图像到输出光流预测结果的时间分别为 0.211 3 s 和 0.119 4 s，由此可见，研究基于自适应光流估计的微表情识别模型的必要性。

此外，本文在复合数据集和 CASME II 数据集上进行了可视化分析。图 4 展示了 AdaMER 模型分别在复合数据集合和 CASME II 数据集上获得的混淆矩阵。对于复合数据集，3 种类别的识别准确率之间存在较少差距，因为复合数据集的样本量较多，3 种类别的样本易于区分。对于 CASME II 数据集，本文的模型在惊讶和高兴上表现出较好的识别性能，这主要得益于本文的模型能捕捉到眉毛和嘴角的运动；但是也存在压抑样本被错误识别为高兴样本的现象，这主要是因为压抑和高兴涉及相似的面部运动单元（AU12、AU15、AU14 和 AU17）。



(a) 复合数据集上的混淆矩阵图　　(b) CASMEII 上的混淆矩阵图

图 4　混淆矩阵可视化

Fig. 4　A visualization of confusion matrix

((a) confusion matrix on the composite dataset; (b) confusion matrix on CASME II dataset)

3.4　消融实验

为了验证提出模型的有效性，本文在 CASME II 单数据集上进行消融实验，结果如表 2 所示，其中，DSDE 表示密集差分编码器，AI、DI 分别表示输入图像为顶点帧和帧间差分图像，OI+AI 代表包含运动信息的两帧。从表 2 可以看出，方法 1 中，顶点帧经过 ResNet18 提取特征进行实验，识别准确率比较低，这也证实了仅从空间上提取特征难以捕捉微表情的细微变化。方法 2 以差分图作为 ResNet18 输入，其准确率比顶点帧时提高了 7.85%，说明包含时序上运动信息对微表情识别准确率的提高有帮助，这也启发了本文基于起始帧和顶点帧设计性能更好的微表情识别模型。

表 2　消融实验结果

Table 2　Results of ablation experiments

方法	输入	DSDE	光流估计	特征融合	ACC	F1
1	AI	×	×	×	0.533 1	0.470 8
2	DI	×	×	×	0.611 6	0.613 0
3	OI+AI	×	×	×	0.739 6	0.723 8
4	OI+AI	√	√	×	0.768 6	0.742 9
5	OI+AI	√	√	√	**0.772 7**	**0.777 2**

注：“√”表示使用该模块，“×”表示未使用该模块。加粗字体表示各列最优结果。

方法 3—方法 5 使用起始帧和顶点帧为主的包含运动信息的两幅图像作为网络输入，用来证明网

络结构的改进是有效的。方法3中使用ResNet18和MEM组成的密集差分编码器输出的语义特征作为分类特征，识别准确率和F1分别达到73.96%和72.38%，实现了比单独使用顶点帧和差分图更优的结果，这表明密集差分编码器能有效提取起始帧和顶点帧之间的运动信息。方法4中由密集差分编码器—解码器构建光流估计任务，编码器输出的特征进行表情预测，与方法3相比，识别准确率和F1分别提升了2.9%和1.91%，这可以证明利用自适应光流估计能有效地捕捉微表情相关的运动特征。在方法5中，将密集差分编码器提取的特征和视觉Transformer提取的光流中判别特征进行融合，相比方法4，识别准确率提升0.41%，F1提升了3.43%，证明特征融合的方式为微表情识别提供面部静态和动态的微表情面部运动特征，使得预测的结果更加可靠。

为了分析优化目标中超参数的敏感性，图5和图6分别给出了参数λ_1和参数λ_2的参数可视化分析，由图5可知，参数λ_1从0.4增加到0.7，F1和准确率均呈现上升趋势，在$\lambda_1=0.7$时获得最佳性能。当$\lambda_1>0.7$时，λ_1的增加会导致性能呈现下降趋势。同理，由图6可知，F1和准确率均随λ_2变化呈波动变化，

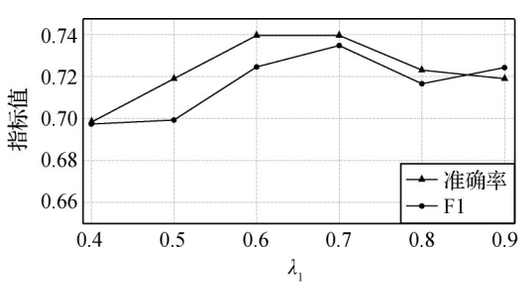


图5　不同λ_1下模型性能变化曲线

Fig. 5　Curves of model performance variation in terms of different λ_1

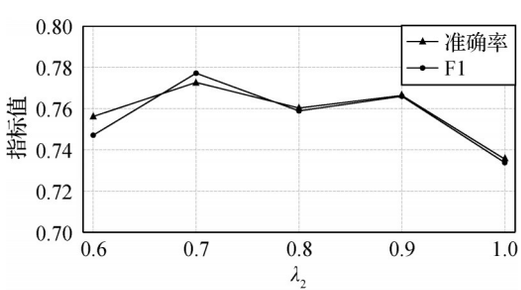


图6　不同λ_2下模型性能变化曲线

Fig. 6　Curves of model performance variation in terms of different λ_2

在$\lambda_2<0.7$时性能呈上升趋势，在$\lambda_2>0.7$时呈下降趋势。因此，本文默认λ_1和λ_2的值均取0.7。

3.5　模型复杂度分析

本节对提出的AdaMER模型进行计算量和参数量分析和比较其中，计算量用每秒浮点运算次数（floating-point operations per second，FLOPs）指标进行衡量。具体地，本文分别对提出AdaMER的光流预测部分与当前主要光流预测模型以及AdaMER微表情推断部分与主流微表情识别模型进行定量比较。

表3列出了AdaMER模型光流预测模块和具有代表性的光流估计方法的模型计算量（FLOPs）和参数量两个指标的值。与Flownet2（Ilg等，2017）相比，本文的光流估计方法显著降低了FLOPs和参数量，能减少推理时间和内存需要，更适合实际应用。虽然本文的光流估计方法的参数量略高于RAFT（recurrent all-pairs field transforms）（Teed和Deng，2020）和GMA（global motion aggregation）（Jiang等，2021），但是FLOPs远低于两者，这主要得益于密集差分编码器中的MEM利用差分信息和多尺度位移信息来捕捉细微的运动变化，而不是计算两帧之间的相关性和多次迭代优化学习的帧间位移。

表3　光流估计方法模型复杂度对比

Table 3　Model complexity comparison of optical flow estimation method

/M

方法	FLOPs	参数量
Flownet2（Ilg等，2017）	50 431.67	162.52
RAFT（Teed和Deng，2020）	39 623.79	5.26
GMA（Jiang等，2021）	44 427.84	5.80
AdaMER（本文）	6 442.60	7.35

表4比较了AdaMER微表情预测部分和基于光流的微表情识别深度学习模型的计算量（FLOPs）、参数量以及在CASME II数据集上的准确率。相比于DSSN（Khor等，2019）、KFC-MER（Su等，2021），AdaMER的计算量FLOPs明显增加，这是由于DSSN和KFC-MER只包含微表情识别部分，而光流需进行离线预提取。与之不同，所提AdaMER直接自适应估计光流并进行微表情识别，同时包含光流提取及微表情识别两部分。值得注意的是，本文提出的

AdaMER 模型在 CASME II 数据集上的识别准确率分别比 DSSN 和 KFC-MER 提高了 6.49% 和 4.51%，这表明自适应光流估计对微表情识别的有效性。此外，与基于 Transformer 提取特征的 SLSTT 模型相比，AdaMER 模型的 FLOPs 指标和参数量分别减少了约 35% 和 60%，但是识别准确率却提高了 1.47%，证明本文所提出的 AdaMER 具有优越性。

表 4　微表情识别方法模型复杂度对比

Table 4　Model complexity comparison of micro-expression recognition

方法	FLOPs/M	参数量/M	准确率/%
DSSN（Khor 等，2019）	667.04	1.26	70.78
KFC-MER（Su 等，2021）	3 798.65	53.24	72.76
SLSTT（Zhang 等，2022）	185 566.86	92.61	75.80
AdaMER（本文）	12 102.50	36.10	**77.27**

注：加粗字体表示最优结果。

4　结　论

针对现有微表情识别方法依赖离线光流估计导致特征表征能力不足问题，本文提出一种端到端的面部运动光流估计与微表情识别多任务学习统一模型 AdaMER。该模型由密集差分编码器—解码器和视觉 Transformer 构成，前者通过运动增强模块实现不同尺度面部差分信息的交互和聚合，增强微表情细微运动感知能力；后者利用注意力机制学习判别性微表情特征。在由 CASME II、SMIC 和 SAMM 构成的复合数据集上进行大量实验和比较，实验结果显示本文提出方法 UF1 和 UAR 分别为 82.89% 和 85.59%，达到了当前最优微表情识别性能，证明了本文所提方法的有效性。然而，本文所提方法模型参数量相对增加，造成训练成本提高。后续工作将探索微表情模型轻量化和压缩技术，减小模型规模和训练开销。

参考文献（References）

Alkaddour M, Tariq U and Dhall A. 2022. Self-supervised approach for facial movement based optical flow. IEEE Transactions on Affective Computing, 13(4): 2071-2085 [DOI: 10.1109/TAFFC.2022.3197622]

Ben X Y, Ren Y, Zhang J P, Wang S J, Kpalma K, Meng W X and Liu Y J. 2022. Video-based facial micro-expression analysis: a survey of datasets, features and algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9): 5826-5846 [DOI: 10.1109/TPAMI.2021.3067464]

Bhushan B. 2015. Study of facial micro-expressions in psychology//Mandal M K, Awasthi A, eds. Understanding Facial Expressions in Communication. New Delhi: Springer: 265-286

Chen B, Liu K H, Xu Y, Wu Q Q and Yao J F. 2023. Block division convolutional network with implicit deep features augmentation for micro-expression recognition. IEEE Transactions on Multimedia, 25: 1345-1358 [DOI: 10.1109/TMM.2022.3141616]

Chen G, Zhang S Q and Zhao X M. 2022. Video sequence-based human facial expression recognition using Transformer networks. Journal of Ilmage and Graphicts, 27(10): 3022-3030 (陈港, 张石清, 赵小明. 2022. 采用 Transformer 网络的视频序列表情识别. 中国图象图形学报, 27(10): 3022-3030 [DOI: 10.11834/jig.210248]

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]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: ICLR

Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, Van Der Smagt P, Cremers D and Brox T. 2015. FlowNet: learning optical flow with convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2758-2766 [DOI: 10.1109/ICCV.2015.316]

Ekman P and Friesen W V. 1975. Unmasking the Face: a Guide to Recognizing Emotions from Facial Clues. Englewood Cliffs: Prentice-Hall: 430-440

Fan X Q, Chen X L, Jiang M J, Shahid A R and Yan H. 2023. SelfME: self-supervised motion learning for micro-expression recognition//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 13834-13843 [DOI: 10.1109/CVPR52729.2023.01329]

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]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A and Brox T. 2017. FlowNet 2.0: evolution of optical flow estimation with deep net-

works//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1647-1655 [DOI: 10.1109/CVPR.2017.179]

Jiang S H, Campbell D, Lu Y, Li H D and Hartley R. 2021. Learning to estimate hidden motions with global motion aggregation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9752-9761 [DOI: 10.1109/ICCV48922.2021.00963]

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. Taipei, China: IEEE: 36-40 [DOI: 10.1109/ICIP.2019.8802965]

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 G H, Yuan Y F, Ben X Y and Zhang J P. 2020. Spatiotemporal attention network for microexpression recognition. Journal of Image and Graphics, 25(11): 2380-2390 (李国豪, 袁一帆, 贲晛烨, 张军平. 2020. 采用时空注意力机制的人脸微表情识别. 中国图象图形学报, 25(11): 2380-2390) [DOI: 10.11834/jig.200325]

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. Shanghai, China: IEEE: 1-6 [DOI: 10.1109/FG.2013.6553717]

Li Y T, Huang X H and Zhao G Y. 2018. Can micro-expression be recognized based on single apex frame?//Proceedings of the 25th IEEE International Conference on Image Processing. Athens, Greece: IEEE: 3094-3098 [DOI: 10.1109/ICIP.2018.8451376]

Li Y T, Huang X H and Zhao G Y. 2021. Joint local and global information learning with single apex frame detection for micro-expression recognition. IEEE Transactions on Image Processing, 30: 249-263 [DOI: 10.1109/TIP.2020.3035042]

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]

Nguyen X B, Duong C N, Li X, Gauch S, Seo H S and Luu K. 2023. Micron-BERT: BERT-based facial micro-expression recognition//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1482-1492 [DOI: 10.1109/CVPR52729.2023.00149]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

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 & Gesture Recognition. Lille, France: IEEE: 1-5 [DOI: 10.1109/FG.2019.8756611]

Shreve M, Godavarthy S, Goldgof D and Sarkar S. 2011. Macro-and micro-expression spotting in long videos using spatio-temporal strain//Proceedings of 2011 IEEE International Conference on Automatic Face & Gesture Recognition. Santa Barbara, USA: IEEE: 51-56 [DOI: 10.1109/FG.2011.5771451].

Su Y T, Zhang J Q, Liu J and Zhai G T. 2021. Key facial components guided micro-expression recognition based on first & second-order motion//Proceedings of 2021 IEEE International Conference on Multimedia and Expo. Shenzhen, China: IEEE: 1-6 [DOI: 10.1109/ICME51207.2021.9428407]

Sun B, Cao S M, Li D L, He J and Yu L J. 2022. Dynamic micro-expression recognition using knowledge distillation. IEEE Transactions on Affective Computing, 13(2): 1037-1043 [DOI: 10.1109/TAFFC.2020.2986962]

Szajnberg N M. 2022. What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS). Journal of the American Psychoanalytic Association, 70(3): 591-595 [DOI: 10.1177/00030651221107681]

Teed Z and Deng J. 2020. RAFT: recurrent all-pairs field transforms for optical flow//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 402-419 [DOI: 10.1007/978-3-030-58536-5_24]

Van Quang N, Chun J and Tokuyama T. 2019. CapsuleNet for micro-expression recognition//Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition. Lille, France: IEEE: 1-7 [DOI: 10.1109/FG.2019.8756544]

Wang F P, Li J, Qi C, Wang L and Wang P. 2023. Multi-scale multimodal micro-expression recognition algorithm based on Transformer [EB/OL]. [2023-07-08]. https://arxiv.org/pdf/2301.02969.pdf

Wang Y D, See J, Phan R C W and Oh Y H. 2014. LBP with six intersection points: reducing redundant information in LBP-TOP for micro-expression recognition//Proceedings of the 12th Asian Conference on Computer Vision. Singapore, Singapore: Springer: 525-537 [DOI: 10.1007/978-3-319-16865-4_34]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

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]

Yan W J, Li X B, Wang S J, Zhao G Y, Liu Y J, Chen Y H and Fu X L. 2014. CASME II: an improved spontaneous micro-expression database and the baseline evaluation. PLoS One, 9(1): #e86041 [DOI: 10.1371/journal.pone.0086041]

Yu J H, Zhang C Y, Song Y and Cai W D. 2021. ICE-GAN: identity-aware and capsule-enhanced GAN with graph-based reasoning for micro-expression recognition and synthesis//Proceedings of 2021 International Joint Conference on Neural Networks. Shenzhen, China: IEEE: 1-8 [DOI: 10.1109/IJCNN52387.2021.9533988]

Zach C, Pock T and Bischof H. 2007. A duality based approach for real-time TV-*L* 1 optical flow//Proceedings of the 29th Joint Pattern Recognition Symposium. Heidelberg, Germany: Springer: 214-223 [DOI: 10.1007/978-3-540-74936-3_22]

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]

Zhang L F, Hong X P, Arandjelović O and Zhao G Y. 2022. Short and long range relation based spatio-temporal Transformer for micro-expression recognition. IEEE Transactions on Affective Computing, 13(4): 1973-1985 [DOI: 10.1109/TAFFC.2022.3213509]

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]

Zheng Z Q, Jiang Q S and Wang S F. 2020. Posed and spontaneous expression distinction through multi-task and adversarial learning. Journal of Image and Graphics, 25(11): 2370-2379 (郑壮强, 姜其胜, 王上飞. 2020. 多任务学习和对抗学习结合的自发与非自发表情识别. 中国图象图形学报, 25(11): 2370-2379) [DOI: 10.11834/jig.200264]

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]

Zhou L, Mao Q R and Xue L Y. 2019. Dual-inception network for cross-database micro-expression recognition//Proceedings of the 14th IEEE International Conference on Automatic Face and Gesture Recognition. Lille, France: IEEE: 1-5 [DOI: 10.1109/FG. 2019. 8756579]

作者简介

包永堂，男，副教授，主要研究方向为机器学习、三维几何建模和虚拟现实。E-mail: baozi0221@sdust.edu.cn

张鹏，通信作者，男，副教授，主要研究方向为计算机视觉和机器学习。E-mail: pengzhang_skd@sdust.edu.cn

武晨曦，女，硕士研究生，主要研究方向为计算机视觉和微表情识别。E-mail: chenxi_wu@sdust.edu.cn

单彩峰，男，教授，主要研究方向为计算机视觉、模式识别和医学图像处分析。E-mail: caifeng.shan@sdust.edu.cn