人脸表情识别可解释性研究综述

张淼萱 张洪刚
（北京邮电大学人工智能学院 北京 100876）

摘 要 近年来，人脸表情识别（Facial Expression Recognition，FER）被广泛应用于医疗、社交机器人、通信、安全等诸多领域.与此同时，为加深研究者对模型本质的认识，确保模型的公平性、隐私保护性与鲁棒性，越来越多的研究者关注表情识别可解释性的研究.本文依据结果可解释、机理可解释、模型可解释的分类原则，对表情识别中的可解释性研究方法进行了分类与总结.具体而言，结果可解释表情识别主要包括基于文本描述和人脸基本结构的方法.机理可解释方法主要研究了表情识别中的注意力机制，以及基于特征解耦和概念学习方法的可解释方法.模型可解释方法主要探究了可解释性分类方法.最后，对表情识别可解释性研究进行了对比与分析，并对未来的发展方向进行了讨论与展望，包括复杂表情的可解释性、多模态情绪识别的可解释性、大模型表情与情绪识别的可解释性以及基于可解释性提升泛化能力四个方面.本文旨在为感兴趣的研究人员提供人脸表情识别可解释性问题研究现状的整理与分析，推动该领域的进一步发展.

关键词 人脸表情识别；可解释性；计算机视觉；情感计算；机器学习
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2024.02819

A Survey on Interpretability of Facial Expression Recognition

ZHANG Miao-Xuan ZHANG Hong-Gang
（*School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing* 100876）

Abstract In recent years, Facial Expression Recognition (FER) has been widely used in medicine, social robotics, communication, security and many other fields. A growing number of researchers are showing interest in the FER area and have proposed useful algorithms. At the same time, the study of FER interpretability has attracted increasing attention from researchers, as it can deepen their understanding of the models and ensure fairness, privacy preservation, and robustness. In this paper, we summarized the interpretability works in the field of FER based on the classification of result interpretability, mechanism interpretability, and model interpretability. Result interpretability indicates the extent to which people with specific experience can consistently understand the results of the models. Specifically, result interpretable FER mainly includes methods based on text description and the basic structure of the face. Wherein the methods based on face structure consists of approaches based on facial action units (AU), topological modeling, caricature images and interference analysis. In addition, mechanism interpretability focuses on explanation of the internal mechanism of the models, including the attention mechanism in FER, as well as the interpretability methods based on feature decoupling and concept learning. As for model interpretability, researchers often try to find out the decision principle or rules of the models. This paper illustrates the interpretable classification methods in

收稿日期：2024-01-23；在线发布日期：2024-09-18. 本课题得到国家自然科学基金面上项目（No. 62076034）资助.**张淼萱**，博士研究生，中国计算机学会（CCF）会员，主要研究领域为情感计算、表情识别、计算机视觉和视觉语言模型等. E-mail：zhangmiaoxuan@bupt.edu.cn. **张洪刚**（通信作者），博士，副教授，博士生导师，中国计算机学会（CCF）会员，主要研究领域为图像检索、计算机视觉和模式识别. E-mail：zhhg@bupt.edu.cn.

FER, which belong to model interpretability. Such approaches involve those based on Multi-Kernel Support Vector Machine (MKSVM) and those based on decision trees and deep forest. Additionally, we compared and analyzed the FER interpretability works. We also identified current problems in this area, including the lack of evaluation metrics for FER interpretability analysis, the challenge of balancing the accuracy and interpretability of FER models, and the limited interpretability data available for expression recognition. Afterwards, a discussion and outlook on the way forward took place. First is about the interpretability of complex expressions recognition, mainly focusing on the compound expressions and more delicate fine-grained expressions. Then it comes to the interpretability of multi-modal emotion recognition. Multi-modal models can obtain better performance by complementing the information of each modality, and their interpretability analysis is also an important direction worth exploring in the future. Additionally, we believe that interpretability of expression and emotion recognition with large models is another significant future direction, including interpretability of Large Vision Models, Vision Language Models and Multi-modal Large Models. Interpretability study can help to improve the safety and reliability of large models. Finally, we address the enhancement of generalization ability based on interpretability. When the models are learning "relevance" rather than "causality", they are easy to make wrong judgments when encountering new data or being affected by other factors, that is, the models do not have good generalization performance. The interpretability analysis helps deepen our understanding of the nature of the models, explain the causal relationship between input and output, and therefore improve the generalization performance. This paper intends to provide interested researchers with a comprehensive review and analysis of the current state of research on the interpretability of facial expression recognition, thereby promoting further advancements in this field.

Keywords facial expression recognition; interpretability; computer vision; affective computing; machine learning

1 引 言

人脸表情识别是情感计算与计算机视觉领域相结合的课题之一,在医疗、社交机器人、通信、安全等各种场景中具有广阔的应用前景和独有的现实意义.随着人脸表情数据集的丰富和计算机计算能力的加强,表情识别算法也获得了不断的发展[1-3].值得注意的是,研究者在关注模型准确率的同时,也越来越多地关注模型的可解释性.可解释性刻画了人类对模型决策或预测过程与结果的理解程度.对于可解释性的研究,有利于说明模型输入到输出之间的因果关系,加深研究者对于模型本质的认识,确保模型的公平性、鲁棒性、隐私保护性能.

目前机器学习中对可解释性并没有统一明确的定义[4],但依据可解释研究的不同视角与面向对象,可解释性研究可以分为三种主要的类型:结果可解释性(result interpretability)、机理可解释性(mechanism interpretability)与模型可解释性(model interpretability).

结果可解释性表明了拥有特定经验的人群能够一致地理解与预测模型结果的程度[5].换言之,如果模型结果是为用户决策使用,则需要根据用户的知识结构对模型结果进行解释.常见的结果可解释性模型包括Kim等人[5]在贝叶斯模型批评框架的启发下,提出的MMD-critic模型,该模型通过有效地学习原型和批评,帮助人类理解.Adler等人[6]利用控制变量法,通过删除某一个特征观察模型的变化,解释模型结果的影响因素.Karpathy等人[7]利用神经网络生成的有关图像的描述性语言来解释网络如何分析图像,具体而言,作者结合了双向递归神经网络以及卷积神经网络以获得双模态嵌入,从而习得图像特征与对应文本.

机理可解释性是一种更高层次的可解释性,是对于模型作用的内部机理的解释,即对模型"透明度"的分析[8].一些机器学习模型的机理是清晰的,

并且可以通过严密的数学推导来证明. 但对于神经网络而言，研究者并不能完全掌握其内部机理. 当前神经网络仍然以试错为主，而对于内部原理和网络具体的学习内容不甚明晰. 而机理可解释性研究即希望通过一些数学工具或算法尝试分析网络内部的工作机理. 常用的可解释性研究方法中，注意力机制、特征解耦和概念学习方法均属于这一类. 注意力机制可以得到不同区域与特征对于模型学习的重要性程度：Zhao等人[9]提出了一种多样化的视觉注意力网络(Diversified Visual Attention Network，DVAN)来解决细粒度对象分类的问题，与无注意力模型相比，该网络显著减轻了对强监督信息的依赖，以学习定位有区别的区域；Zeng等人[10]介绍了一种基于双记忆注意力的方面级别语句情感分类模型，此模型借助循环神经网络的序列学习能力得到语句编码，同时构造相应的注意力机制从语句编码中提取出关于给定方面词的情感表达. 特征解耦方法则是通过解耦方法，减小各特征维度之间的相关性，分离出与任务相关的人类可理解特征：如Feng等人[11]通过自监督的学习方法，将旋转特征辨别与实例特征辨别解耦，通过减轻旋转标签噪声的影响来改进旋转预测，以将旋转不变性结合到特征学习框架中. Zhou等人[12]提出类引导特征解耦网络(Class-Guided Feature Decoupling Network，CGFDN)用于航空图像分割，通过设计特征解耦模块，将不同类别对象之间的同现关系编码为卷积特征. 概念学习的方法则往往基于对特定概念的选择和学习，利用概念加深对模型的内部特征或机理的理解. 例如Yang等人[13]使用语义激活张量(Semantic Activation Tensors，SAT)表征语义概念，通过将其与网络梯度的内积来量化语义对分类结果的重要程度，从而提高卷积神经网络的可解释性.

模型可解释性则针对模型决策原理进行解释[14]. 这一类的典型可解释性方法包括决策树和规则学习等. 在决策树方法中，人们利用树状结构来表示决策问题中的步骤、条件、方案、损益值、结果等. 例如Krishnan等人[15]提出了一种从已训练神经网络生成的输入数据中提取决策树的方法，提取的决策树除了执行分类之外，还可以用于理解神经网络的工作；Yang等人[16]使用一个从贡献矩阵中习得的紧凑二叉树——解释树，来显式地表示黑匣子机器学习模型中隐含的最重要的决策规则. 规则学习则是从训练数据中习得一组可对示例进行判断的规则，可以使用分解法或教学法提取规则[17]. 分解法一般特定于模型，通过对底层预测模型内部结构的分析与分解提取规则. 例如文献[18]，通过使用三阶段算法提取规则来理解神经网络：首先构建一个权重衰减反向传播网络，然后对网络进行修剪，最后通过递归离散隐藏单元激活值来提取规则；教学法则是在模型不可知的情况下，将底层预测模型视为黑盒，并使用所提供的输入和输出之间的关系理解模型，这往往可以通过利用一个新的、更简单、可解释性也更强的模型来逼近原始模型的输出来实现. 如文献[19]，提出了HYPINV方法，以超平面的形式提取与神经网络具有相似输入输出关系的规则体系.

本文主要针对人脸表情识别的可解释研究进行整理，旨在帮助相关领域的研究人员更全面地了解人脸表情识别可解释性问题的研究现状，推动该领域的进一步发展. 如图1所示，文中按照上述可解释性分类方法对表情识别可解释性研究进行了分类，并以表情识别过程自前至后的顺序对相关研究进行整理总结. 首先是针对用于表情识别的图片本身进行解释，其中基于文本描述的方法对图片整体进行文字解释，而基于人脸基本结构的方法定位和利用图片中与表情相关的部分. 然后进一步考虑特征层面的解释，使用注意力机制的方法通过学习特征的重要程度进行解释，特征解耦与概念学习方法则尝试将高维、纠缠的特征转化为独立、可理解的. 最后是对表情识别模型的分类与决策过程的解释，即可解释性分类方法. 而按照可解释性研究的分类方法，文中整理的基于文本描述和人脸基本结构的方法，属于结果可解释方法，这里主要是在人脸表情的图片层面使人类可理解表情识别结果. 文本描述方法通过自然语言模块，将模型识别结果与原理转化成人类可理解的描述形式. 而基于人脸基本结构的方法又包括基于面部动作单元(Action Unit，AU)[20]、基于拓扑建模、基于表情图片漫画化和基于干扰分析的方法. 这些方法利用人脸AU或语义拓扑图与面部区域和肌肉单元的联系，或利用漫画对面部动作的夸张，使模型定位到与表情相关的人脸具体区域及其变化，或利用干扰的影响找到对模型结果重要的面部区域，从而达到解释模型的效果，同样属于结果可解释方法. 然后是基于注意力机制的可解释表情识别方法，属于机理可解释性方法，通过注意力计算对模型内部机理进行分析. 之后是基于特征解耦与概念学习的可解释方法，将特征转化为更加清晰、易于理解的形式，同样可以归为机理可解释性. 最后整理了可解释性分类方法，主要对表

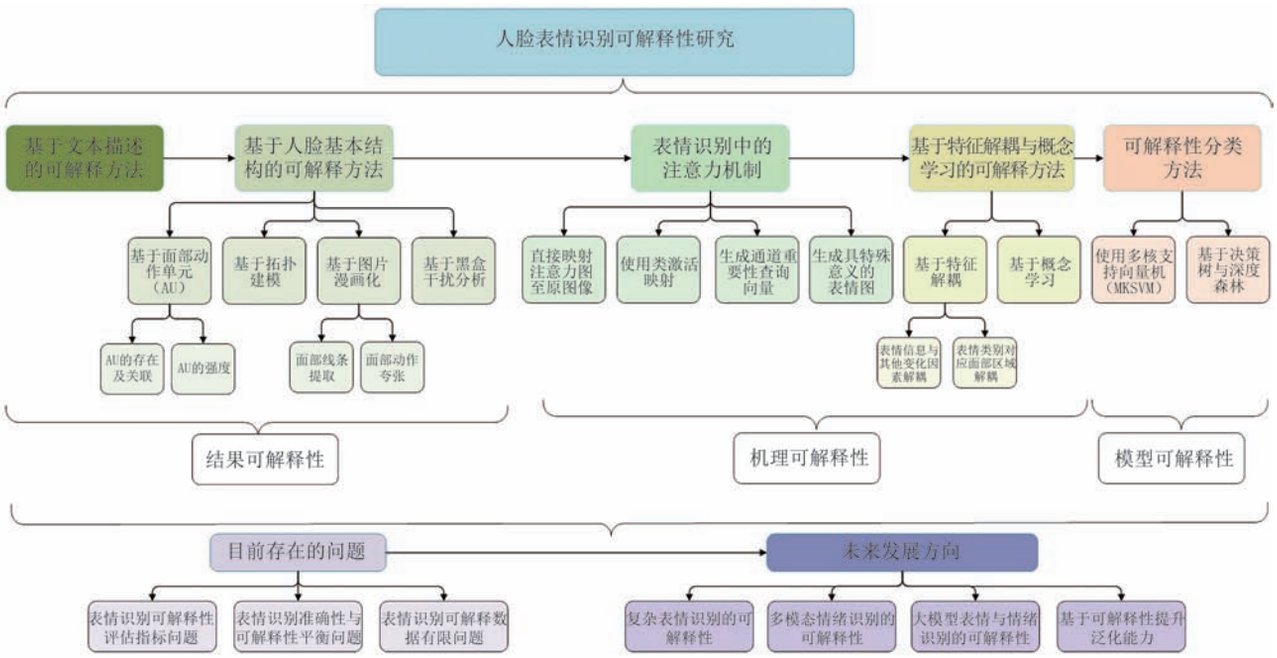


图1 全文整体结构

情识别模型的分类和决策过程进行解释，属于模型可解释方法，包括使用多核支持向量机[21]的方法，以及基于决策树与深度森林的方法.

接着，我们对可解释性表情识别方法的实验结果进行了比较与分析，并意识到，尽管表情识别可解释性研究已经取得了一些可喜的成就，但目前的研究仍然存在着不确切与不充分之处，具有较大的发展空间. 本文针对这一方向目前存在的问题进行了探索，包括表情识别可解释性评估指标问题、表情识别准确性与可解释性平衡问题，以及表情识别可解释数据有限问题. 最后，本文预测了表情识别可解释性这一领域未来可能的研究方向. 首先是复杂表情识别的可解释性，包括复合表情和更精细的细粒度表情. 然后是多模态情感识别的可解释性，多模态模型可以通过各个模态信息的互补获得更好的效果，其可解释性也是未来值得探索的一个重要方向. 之后是关于大模型表情与情绪识别的可解释性，包括大视觉模型（Large Vision Model，LVM）、视觉语言模型（Vision Language Model，VLM）与多模态大模型的可解释性. 近期内大模型取得了飞速的发展，这也为情感计算领域带来了更多的可能性，有关大模型可解释性的研究，有利于提高大模型的安全性与可信性，具有重要意义. 最后，我们关注基于可解释性的泛化能力的提升. 当模型学习的是“相关性”而非“因果性”时，在遇到新数据或受到其他因素影响时，很容易做出错误的判断，即模型不具备良好的泛化性能. 而可解释性研究将有助于加深研究者对模型本质的理解，解释输入和输出之间的因果关系，从而提高模型的泛化性能.

2 基于文本描述的可解释方法

基于文本描述的方法一般用于存在语言模块的多模态模型中，通过自然语言描述的方式揭示神经网络是如何分析图像的，从而达到解释模型的效果，属于结果可解释方法. 近期内，视觉-语言模型的不断发展，也为基于文本描述的可解释方法注入了新的活力.

在表情识别领域，Li等人[22]提出了一种基于对比文本-图像对预训练方法（Contrastive Language-Image Pre-training，CLIP）[23]的静态、动态表情识别统一框架——CLIPER. CLIPER中使用了多表情文本描述符（Multiple Expression Text Descriptors，METD），它可以为每个表情自动学习一组文本描述符，这些描述符对应于同一表情的不同形式，如图2(c)所示. 实验表明，METD从两方面提高了模型的可解释性：第一，METD所学习的文本描述符具有可解释性. 如表1所示，通过观察其在字典中最接近的词，发现不仅有一些与表情直接相关的文本（如“saddest”表示悲伤，“kindness”表示快乐），还有一些间接相关的语义（如“roaring”代表愤怒和“zombies”

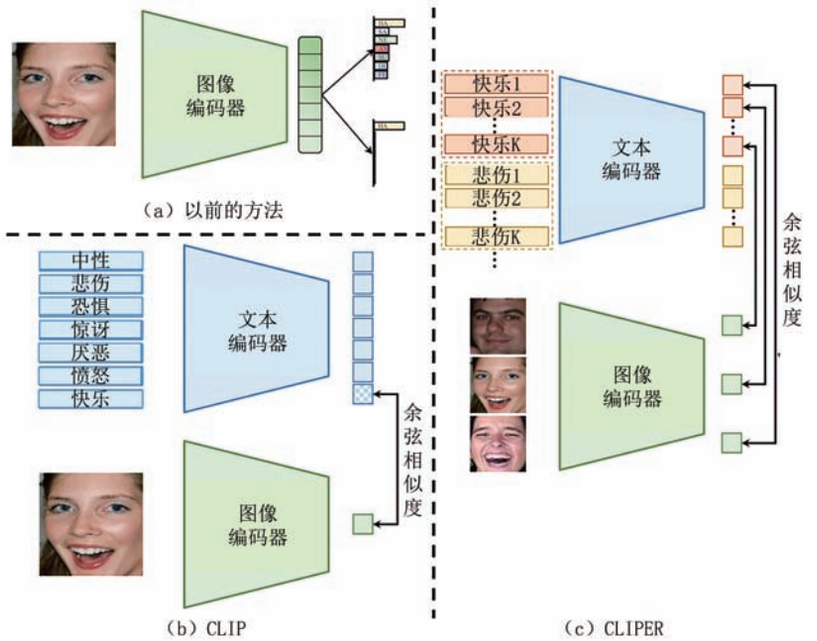


图2　文献[22]中展示的三种表情识别框架((a)是最常用的使用独热标签或软标签训练模型的方法，可解释性较差.(b)是标准CLIP[23]框架，为每个类设计一个文本描述符，并使用图像和文本嵌入之间的余弦相似性进行推断；但由于FER任务标签较为抽象，效果不佳.(c)是CLIPER方法[22])

表1　CLIPER学习的多表达文本描述符的最近单词[22]

表情	索引和最接近的词
惊讶	(2,3) zombies; (3,3) possibly; (4,4) nugget
恐惧	(1,3) "!"; (3,4) ned; (5,3) indictment
厌恶	(2,1) sanctions; (2,2) fern; (4,1) boredom
快乐	(2,2) benefits; (3,1) kindness; (4,2) hopeful
悲伤	(1,1) defeats; (2,3) saddest; (3,3) smear
愤怒	(1,3) berser; (3,1) roaring; (3,2) slam
中性	(2,2) abstraction; (3,2) relaxation; (4,1) loose

注：由于空间的限制，作者只显示了一些最近的单词，并给出它们的对应索引(k,m)，这表示对应表情超类的第k个子类的第m个学习文本标记.

代表惊讶)；第二，METD学习了更细粒度的表情语义.如图3中显示的面部表情及其对应的子类，对于"厌恶"，METD学习了两种不同的表达形式；对于"快乐"，METD根据快乐的强度(微笑或大笑)划分了子类；而对于悲伤，则根据年龄来划分子类.

类似地，Tao等人[24]提出了一个基于CLIP的，在情感、动态和双向三方面实现对齐的动态表情识别的模型A^{3}lign-DFER，其中包括利用可学习的多维对齐令牌(Multi-dimensional Alignment Token，MAT)替换CLIP输入中的标签文本.作者展示了与MAT所学习的标记具有高度相似嵌入的词汇，发现MAT成功学习了与表情相关的单词，且某些学习的嵌入与表情描述完全匹配，这也表现了方法具有较好的可解释性.

图3　文献[22]中几种表情及其对应的利用METD划分的不同子类

而Foteinopoulou等人[25]利用关于上下文、表情或情感线索的文本描述作为自然语言监督，来增强丰富潜在表示的学习，从而实现动态表情的分类.文本描述的引入不仅有利于提高分类性能，也有助于人类研究者理解模型构建决策的影响因素，使模型获得一定程度的可解释性.

3 基于人脸基本结构的可解释方法

本节中主要介绍了基于人脸基本结构的可解释性方法，通过分析与表情识别密切相关的面部结构，达到了识别结果可解释的效果，属于结果可解释方法. 这类方法包括基于面部AU[20]、基于拓扑建模、基于图片漫画化的方法，以及基于面部结构干扰分析的方法. 其中基于AU的方法包括利用AU的存在、关联与强度解释模型；基于拓扑建模的方法主要利用人脸关键点进行拓扑建模，捕获人脸语义信息，提高可解释性；基于图片漫画化的方法则利用面部线条提取和面部动作夸张的方法将真实表情图片漫画化，利用漫画的夸张性，帮助识别和解释过程. 基于干扰分析的方法会通过对面部不同部分添加干扰并对其进行分析从而确定对于模型的重要区域，以便对模型进行可解释分析.

3.1 基于面部动作单元

美国心理学家Ekman Paul和Friesen开发了面部动作编码系统(Facial Action Coding System，FACS)[20]，他们根据人脸的解剖学特点，将其划分成若干既相互独立又相互联系的动作单元(Action Unit，AU)，并分析了这些运动单元的运动特征及其所控制的主要区域以及与之相关的表情. FACS中一共定义了44种面部动作编码，几乎所有可能的表情都可以由不同的面部动作编码组合而成.

通过AU的存在、关联或强度来帮助判断表情种类，将各类表情与AU相联系，使得表情相关的肌肉部位和变化得以被捕捉，提高了表情识别模型的可解释性.

3.1.1 利用AU的存在与关联提高可解释性

AU的存在解释了不同的表情下人脸肌肉和面部动作的变化情况，在帮助模型定位与识别面部表情的同时，也提升了模型的可解释性.

Simon等人[26]提出了一种基于分段的方法kSeg-SVM，综合了静态建模和时间建模，从视频中自动检测AU. Liu等人[27]提出AU感知深度网络(AU-aware Deep Networks，AUDN)，利用三个顺序模块，根据面部AU的解释性来学习特征，从而帮助表情识别. Khorrami等人[28]研究了用于表情识别的卷积神经网络(Convolutional Neural Network，CNN)，并通过实验证明CNN所学习的特征与Ekman等[20]提出的人脸AU非常一致，如图4所示，作者使用去卷积网络在像素空间中显示单个被激活的神经元相应的空间模式，并发现许多滤波器都是由与面部动作单元相对应的区域激活的. Liang等人[29]提出了一种基于端到端多尺度AU的网络(Multi-Scale Action-Unit-based Network，MSAU-Net)，用于图像的面部表情识别，该网络直接关注面部AU定位并利用“放大”操作来聚合不同的局部特征，从而学习更强大的面部表示. 作者探究了AU检测对其表情识别方法的积极影响，并可视化了AU特异性感受野，如图5所示，并据此探究面部最具辨别力的部位在哪里. Tang等人[30]介绍了一种称为PIAP-DF的联合策略以解决AU检测中的一些难点，包括使用对每个AU进行像素级注意力的多阶段像素关注学习方法，来解决细粒度和鲁棒的局部AU信息提取不佳问题. 使用反个人特定特征方法，旨在尽可能消除与任何个人身份相关的特征，以及利用具有离散反馈的半监督学习方法，从而有效利用未标记的数据并减轻错误标签的负面影响.

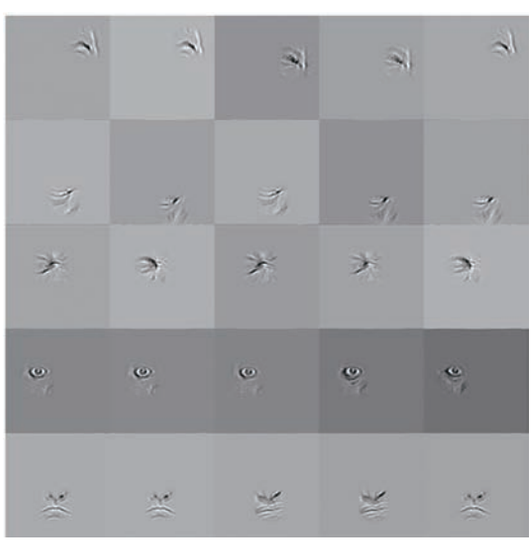


图4 文献[28]中，网络的第三层卷积中，激活五个选定滤波器的面部区域的可视化(每一行对应于conv3层中的一个滤波器)

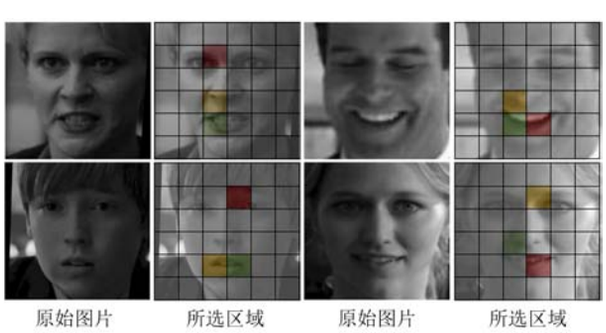


图5 文献[29]提出的框架中，由渐进顺序生成的前三个AU特异性感受野的可视化(分别用红色、绿色、黄色表示)

Shingjergi等人[31]提出了一种方法，使用AU作为解释FER模型操作和结果的手段. 他们利用与FER并行的AU检测，将识别出的AU翻译成自然语言描述，从而构成了人类可理解的解释.

除此之外，越来越多的研究者注意到了对于AUs之间相关关系的利用. 由于生理原因，AUs之间往往彼此依赖，具有强相关关系. 将AUs的相互关系应用于AU检测之中，提高了模型的识别性能与可解释性. Zhao等人[32]提出了一种联合图像块和多标签学习(Joint Patch and Multi-label Learning，JPML)框架，对特征、AU及其相互作用背后的结构化联合依赖性进行建模；具体地，JPML利用群体稀疏性来识别重要面部图像块，并学习受AU共现概率约束的多标签分类器.

近年来，由于Transformer模型[33]在相关关系建模方面卓有成效，研究者尝试利用Transformer模型对AUs之间的相互关系进行建模. Tallec等人[34]设计了一种多标签检测Transformer，利用多头注意力来学习人脸图像中与每个待预测AU最相关的部分. 更准确地说，论文中使用两步来选择对每个AU合适的特征，第一步是在每个已编码的人脸图像块中使用自注意力机制提取特征，第二步是对这些特征和已学习的AU图像块应用交叉注意力机制进一步选择更具价值的特征. Wang等人[35]利用局部时空特征和标签式面部AU相关性，提出了一种基于Transformer的模型. 具体而言，作者设计了一个基于视觉时空Transformer的模型和一个基于卷积的音频模型来提取AU特定特征，同时，受面部AU之间关系的启发，提出了一种基于Transformer的相关模块来学习AU之间的相关性，如图6所示.

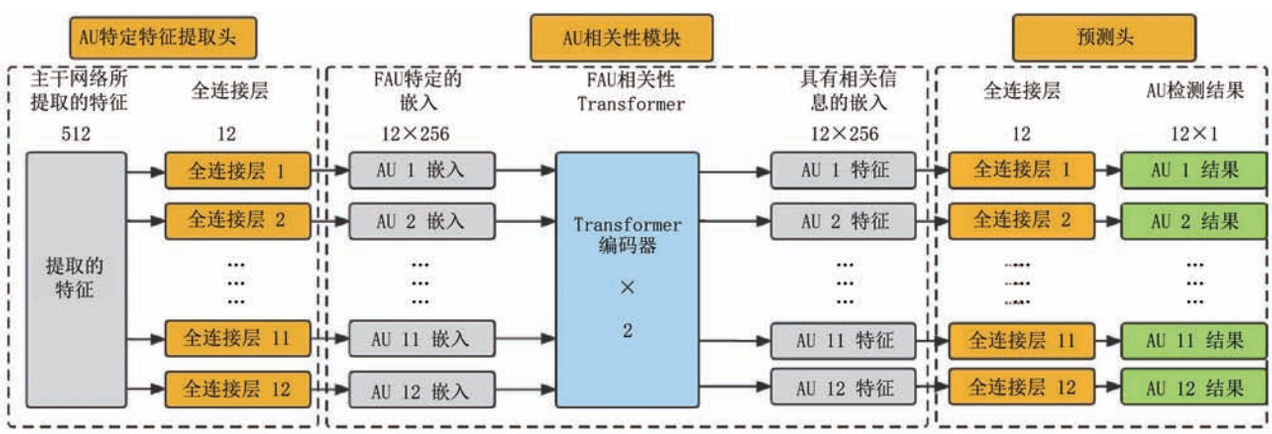


图6　文献[35]中人脸AU相关性学习模块

3.1.2　利用AU的强度提高可解释性

除了基于AU的存在与关联的方法，还有一类模型旨在检测AU的强度，并利用AU及其强度，定位表情相关的面部区域与组件，提升了模型的可解释性.

Li等人[36]提出了一个基于动态贝叶斯网络(Dynamic Bayesian Network，DBN)的框架来系统地建模多级AU强度之间的动态和语义关系，并将提取的图像观测值与所提出的DBN模型系统集成，然后通过概率推理来实现AU强度识别. Benitez-Quiroz等人[37]提出了一种计算机视觉算法，以注释一个包含100万张自然环境面部表情图像的大型数据库，在数据库中可靠地识别AU及其强度，并提出表情识别数据库Emotionet. Walecki等人[38]引入了一个新的建模框架：Copula序数回归，利用Copula函数和条件随机场，从AU强度的边缘建模中确定AU相关性的概率建模，实现了多个AU强度的联合学习和推断，同时具有计算可处理性. Zhang等人[39]提出了一种基于弱监督图像块的，由特征融合模块和标签融合模块组成的深度模型，用于多个AU的联合强度估计，如图7所示，在训练期间，模型使用一个序列作为输入，并使用序列级标签(峰和谷帧的强度注释)来提供监督. 在推理过程中，使用单帧作为输入，并联合学习到的任务相关上下文，以进行上下文感知的AU强度估计. Deramgozin等人[40]提出了一种混合AI可解释框架(Hybrid AI Explainable Framework，HEF)，由一个CNN结构的主功能管道和一个可解释管道组成. 其中可解释管道利用面部动作单元(Facial Action Unit，FAU)强度的捕捉和模型不可知解释方法LIME[41]解释了表情识别的结果，并通过FAU强度结合多层感知机(Multi-Layer Perceptron，MLP)强化主功能管道提供的决策.

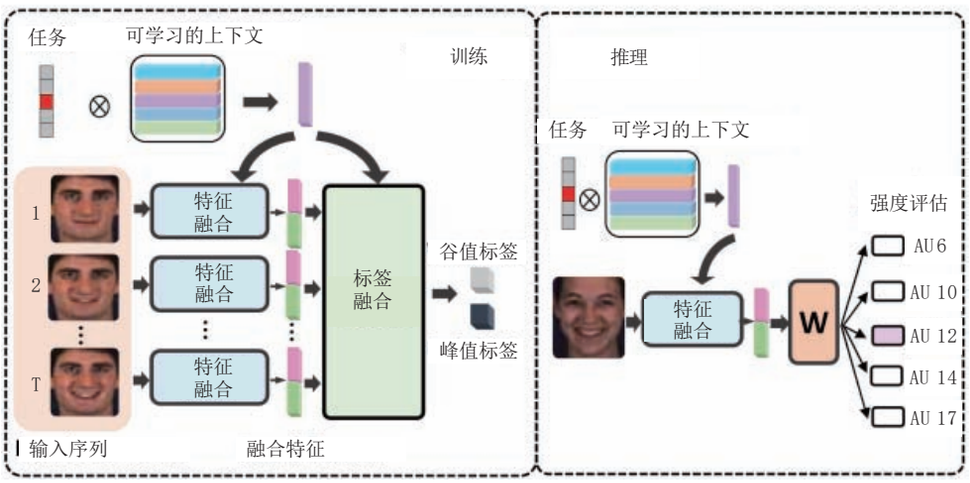


图7 文献[39]所提出方法的训练和推理阶段

3.2 基于拓扑建模

人脸的空间拓扑图是表示人脸的一种常用方法，决定了面部的哪些部分以及要表示的关系．在神经网络中，常常通过人脸关键点的检测和筛选来确定面部结构的拓扑图．通过对面部拓扑结构的学习，在识别人脸表情的同时，也分析出了决定表情的面部组件及关系，即通过对人脸语义信息的探索，提高了表情识别的可解释性．

Zhou等人[42]提出了空间时间语义图网络(Spatial Temporal Semantic Graph Network，STSGN)用于表情识别，STSGN通过面部拓扑结构的端到端特征学习来自动学习空间和时间模式，图8展示了STSGN方法的整体流程．论文提出的语义人脸图将时间动力学和共现动作单元也编码在人脸图中，以便于该方法的解释性．将STSGN学习的内部图形特征可视化，如图9所示，可以直观地链接到原始输入的特定部分，有较强的可解释性．例如对于快乐和惊讶的表情，眉毛抬起和下巴下垂这些动作和区域被明显激活；而对于悲伤表情，则使用皱眉和压唇动作进行识别．类似地，Liu等人[43]也通过识别

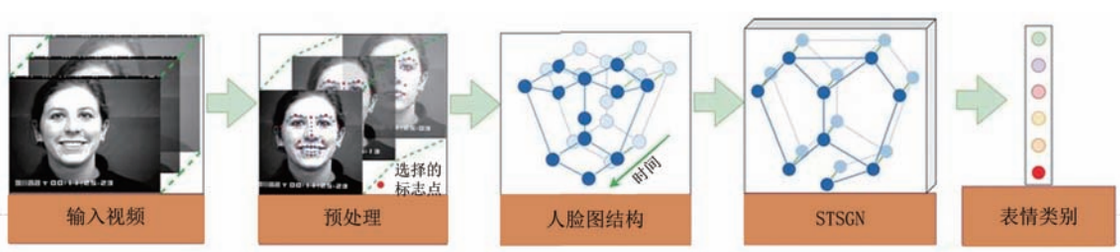


图8 STSGN方法[42]的整体流程

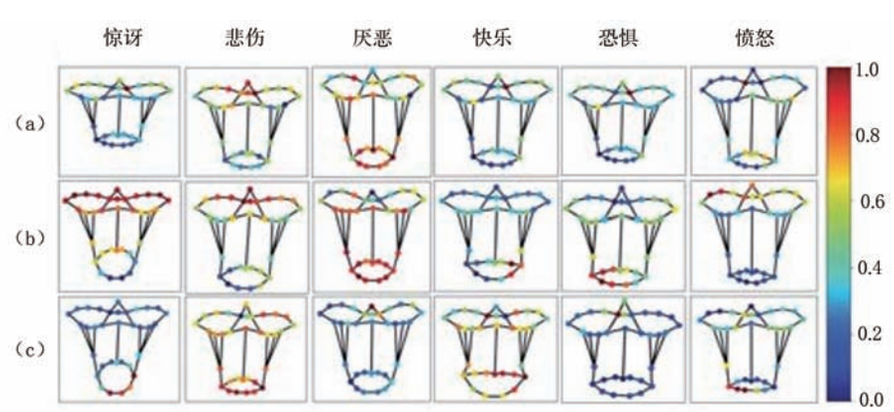


图9 从上到下，显示了STSGN在每个表情的初始(a)、过渡(b)和峰值状态(c)下学习的图形特征图[42](右侧的颜色栏根据每个节点的颜色说明其重要性)

面部拓扑结构提升了面部表情的可解释性. 论文中提出了基于语义图的双流网络(Semantic Graph-based Dual-Stream Network, SG-DSN),该网络设计了一种图表示来建模关键外观和几何面部变化及其语义关系. 为了探索特征的语义,对SG-DSN学习的特征进行可视化,得到的观察结果与生理解剖学和认知神经学中的理论一致,这证明SG-DSN可以显式提取面部表情的语义信息,并具有一定的可解释性.

而Liu等人[44]则提出了一种利用图神经网络(Graph Neural Network, GNN)[45],基于人类视觉认知策略的FER方法,如图10所示,模型根据区域划分机制将检测到的人脸划分为6个区域,然后通过局部视觉识别过程选择每个区域中更具代表性的特征点作为关键特征点,并通过区域协同识别过程将选择的关键节点进行模糊连接,从而基于人脸拓扑结构构建GNN模型. 通过GNN考虑了人脸不同区域关键点之间的关系,提升了模型的可解释性.

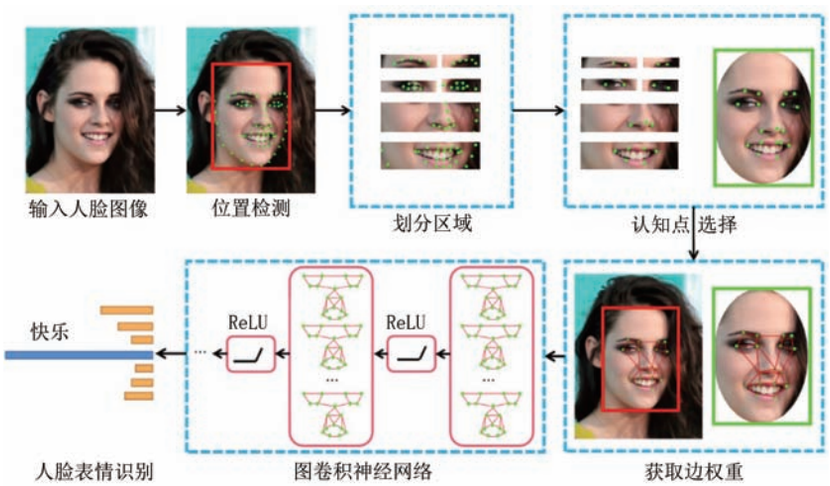


图10　Liu等人提出的基于图神经网络方法的示意图[44]

3.3　基于图片漫画化

人脸漫画是一种基于真实人脸图片,并对特定部分进行提取、夸张从而得到的图片. 漫画形式有利于放大其主体的特征,尤其是那些与众不同,有别于其他的特征[46]. 研究者将漫画的这一特点应用于表情识别过程中,通过将表情相关的面部动作进行提取、放大和夸张化,即将图片"漫画化",以辅助识别. 与此同时,漫画化的图片突出了人脸上与表情相关的独特特征,引导模型对其进行学习,也提升了模型的可解释性.

3.3.1　通过面部线条提取进行漫画化

研究者可以通过特征提取方法提取面部轮廓与动作的线条,生成漫画化的图像,然后根据漫画图像进行识别. 通过面部线条的提取,原图片中与表情信息无关的冗余信息被大量去除,在漫画化的图像中留存了更简洁、与表情信息相关度更高的信息.

Gao等人[47]提出了一种使用基于线条的方法,将输入的静态图片漫画化,然后进行表情识别,该方法使用用户草图表情模型的结构和几何特征来匹配输入面部图像的线边缘图(Line Edge Map, LEM)描述符,并定义了对表情变化鲁棒的视差度量,图11展示了该方法生成的基于线条的漫画.

3.3.2　通过面部动作夸张进行漫画化

在此类方法中,研究者常常通过表情图片中的标志点信息,将面部表情与参考标准(如中性表情)相对比,并夸大表情与其参考之间的差距,从而将图片进行漫画化. 研究表明,模型在识别漫画化的图片时对表情往往有更好的感知能力,证明了图片漫画化能引导模型更多地关注与表情相关的面部区域.

Calder等人[48]通过增大表情与参考标准之间的差距产生了漫画化的图像,通过缩小差距生成了"反漫画",如图12所示;并通过实验验证了:与真实表情相比,漫画图像的识别速度明显更快,而反漫画图像的速度明显较慢;此外,中性表情(面部肌肉放松的状态)和平均表情(6类基本表情的平均)的漫画产生了相同的结果模式. 类似地,Calder等人[49]验证了随着漫画夸张程度的增加,情绪强度也随之增加. Leppänen等人[50]表明情绪分类的速度和准确性随着漫画化程度,即表情强度的增加而提高.

与上述文献不同,Furl等人[51]不是针对静态图片,而是针对视频中的人脸表情进行漫画化. 具

图11 文献[47]中使用LEM描述符生成的中性(a)、微笑(b)和尖叫表情(c)

体而言,作者使用从视频中提取的物理运动指标,在距离正常运动(即平均或原型运动)更远的时空面部空间中的某个位置进行渲染,从而漫画化了面部特征运动的时空信息,包括大小、速度和时间.实验表明,时空漫画增加了不同类表情的感知差异.

3.4 基于黑盒干扰分析

基于干扰分析的方法可以通过分析模型在应对不同的图像扰动时的反应,从而判断图像中对于模型而言的重要区域,从而达到解释模型的目的.这类方法不需要通过内在的模型结构来操纵或观察模型层的输出,只需要黑盒模型的输入输出函数,是真正的黑盒解释方法.

Mery等人[52]提出了一种称为MinPlus的显著性图方法,可用于解释包括表情识别在内的任何面部分析方法,而无须在识别模型内部进行任何操作.MinPlus的关键思想是基于对给定图像被扰动时识别概率如何变化的分析,具体而言,MinPlus删除并聚合图像的不同部分,并单独和协作地测量这些部分的贡献,如图13所示.

在文献中,作者使用MinPlus测试了基于Xception架构[53]的人脸表情识别模型,选择了FER-2013数据集的一些图像,并得到结果如图14所示.

图12 文献[48]中的原始图片(0%)以及基于平均表情(a)与中性表情(b)生成的漫画(+50%)、反漫画(-50%)图片

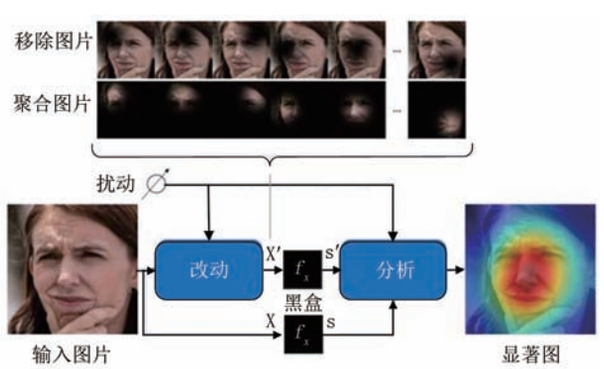


图13 MinPlus方法[52]的示意图(其中"f_x"是黑盒的输入输出函数)

可以看出,MinPlus得到了算法对应的面部重要区域,由此达到了解释模型的效果.同时,作者通过实验表明,与其他一些黑盒解释方法,如LIME[41],RISE[54],RISEguass(使用高斯掩码的RISE算法)[52]等比较,MinPlus提供了更为稳定的显著性映射.

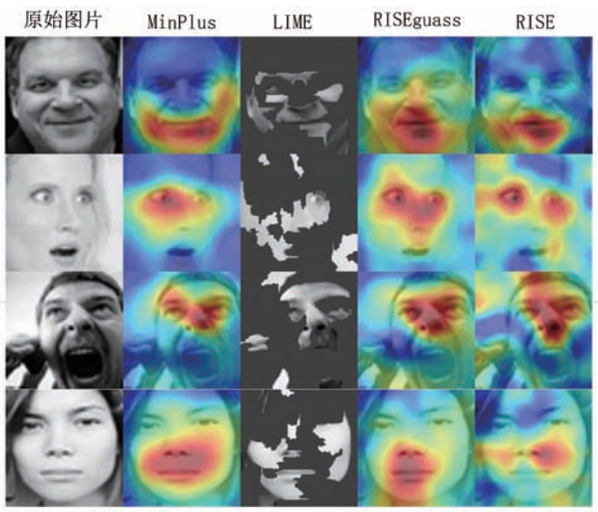


图14　文献[52]中，利用MinPlus[52]以及LIME[41]，RISEguass[52]，RISE[54]获得的使用Xception识别表情的可解释性（图中显示了四种表情的表现，按照行分别是快乐、惊讶、愤怒和中性）

4　表情识别中的注意力机制

由于在面部表情的识别过程中，某些面部区域提供了更具有辨识力的信息，因此，使用注意力机制，给予不同的面部特征不同的权重对于提高识别准确度而言是有益的. 同时，注意力机制的学习可以得到不同面部区域与特征对于模型学习的重要性程度，从而提高模型的可解释性. 基于注意力机制的方法属于机理可解释方法.

本节主要对基于注意力机制的可解释表情识别方法进行总结，而为了使注意力机制所学习的特征重要性能够以人类可以理解的形式展现出来，需要用到一些可视化的方法，下面总结了四类使用注意力解释模型并可视化的方法：直接将调整大小后的注意力图映射到原图像、使用类激活映射[55]、使用注意力机制生成通道重要性查询向量以及生成具有特殊意义的特征图.

4.1　直接映射注意力图至原图像

这类方法在学得注意力图后，直接将生成的注意力图调整到原图像大小，再映射到原图像以进行可视化. Xie等人[56]提出了深度注意力多径卷积神经网络（Deep Attentive Multi-path Convolutional Neural Network，DAM-CNN），通过注意力掩码自适应估计不同图像区域对于面部表情识别的重要性，并将表情信息从无关变化中分离出来. 模型生成了7×7的注意力掩码以表示输入图像中某个区域的重要性，然后通过双线性插值直接调整掩码图的大小至与输入图像匹配，最后将生成的热图映射到原图像中，图15展示了在受限数据集（CK+[57]和JAFFE[58]）和无约束数据集（BAUM-2i[59]、FER2013[60]和SFEW[61]）中，应用DAM-CNN获得的显著表情区域的可视化. Chen等人[62]提出了面部运动先验网络（Facial Motion Prior Networks，FMPN），引入了一个加法分支来生成面部掩码，以便专注于面部肌肉运动区域，并通过使用中性人脸和相应带有表情的人脸之间的平均差异作为训练指导，以指导面部掩码的学习；作者将掩码应用于PFN中的原始输入表情面部并与之融合，获得了可视化结果. Xue等人[63]提出了可以学习丰富的关系感知局部表示的TransFER模型，是第一次将Transformer模型[33]应用于表情识别领域，主要由多注意力下降（Multi-Attention Dropping，MAD）、ViT-FER和多头自注意力下降（Multi-head Self-Attention Dropping，MSAD）三部分组成；作者采用文献[64]的方法进行可视化以解释模型，首先将可视化注意力图调整到与输入图像相同的大小，并通过COL_ORMAP JET颜色映射将注意力图可视化到原始图像.

4.2　使用类激活映射

使用类激活映射（Class Activation Mapping，CAM）[55]也是一种基于注意力机制的可解释表情识

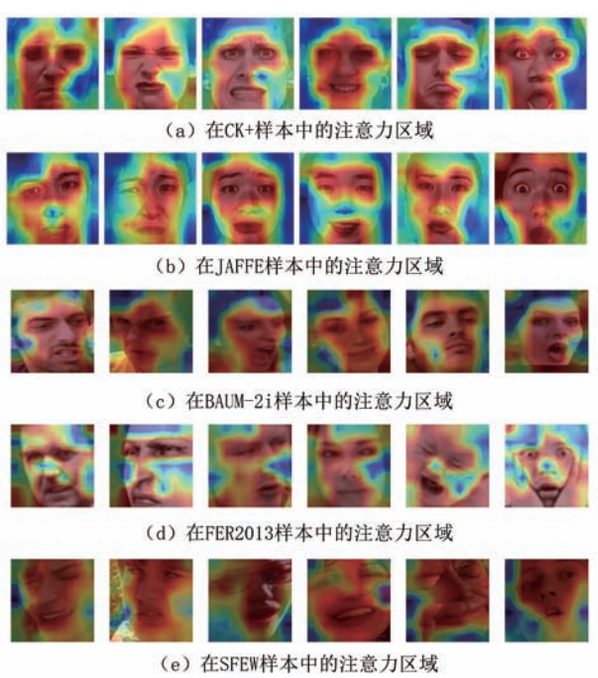


图15　文献[56]展示的在不同数据集中应用DAM-CNN获得的显著表情区域的可视化

别模型可视化方法.CAM是一种生成热力图的技术,可用于可视化卷积神经网络,突出图像的类的特定区域.利用CAM,可以可视化地显示哪些特征是模型所更加关注的,从而提升可解释性.Zhang等人[65]提出了一种对称注意一致性(Erasing Attention Consistency, EAC)方法来自动抑制训练过程中的噪声样本,以解决带噪标签的FER问题;具体而言,作者首先利用人脸图像的对称语义一致性来设计一个不平衡框架,然后随机擦除输入图像,并使用对称注意力一致性来防止模型聚焦于部分特征.论文中使用了CAM提升模型性能,并可视化给定图像的注意力图,以展示模型所关注的特征,如图16所示.当然,CAM与其变体Grad-CAM[66]、Grad-CAM++[67]等通过可视化解释模型,也可实现对模型的评估.Li等人[68]则提出了一种具有注意机制的卷积神经网络(Convolution Neutral Network with Attention mechanism, ACNN),它可以感知人脸的遮挡区域,并聚焦于最具辨别力的未遮挡区域;ACNN组合了来自面部感兴趣区域的多个表示,通过提出的门单元对每个表示进行加权,该门单元根据无障碍性和重要性从区域本身计算自适应权重.论文中采用Grad-CAM[66]进行可视化,如图17、18所示,分别生成了无遮挡图片、合成遮挡图片和真实遮挡图片的注意力图,直观表现出模型于人脸图像上的聚焦情况,评估模型所关注的特征是否合理的同时提高了模型可解释性.Wen等人[69]提出分散注意力网络(Distract your Attention Network, DAN)用于表情识别,其中首先使用特征聚类网络来提取鲁棒特征,然后利用多头注意力网络同时关注多个面部区域,之后设计注意力融合网络分散注意力并融合多个注意力图,如图19所示,作者利用Grad-CAM++[67]可视化地对比了单注意力头的模型和

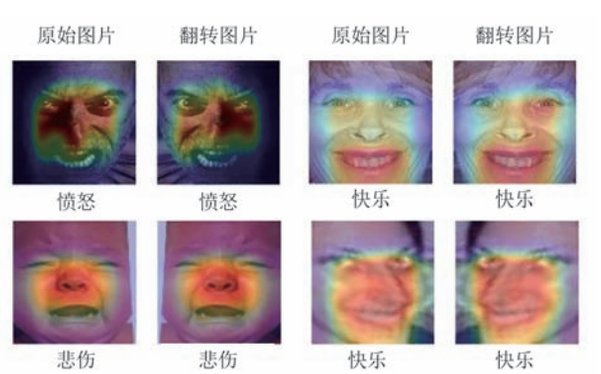


图16 文献[65]中使用CAM生成的EAC方法在原始图像及其翻转对应图像上的注意力图

图17 使用Grad-CAM生成的关于原始图像及其对应的合成遮挡图像的注意力图[68](其中图像的表情标签在其左侧,不同方法的预测结果位于图片上方;CNN代表卷积神经网络,ACNN代表具有注意力机制的CNN[68],'p'代表基于局部图像块,'g'代表集成图像块级别的局部表示和图像级别的全局表示)

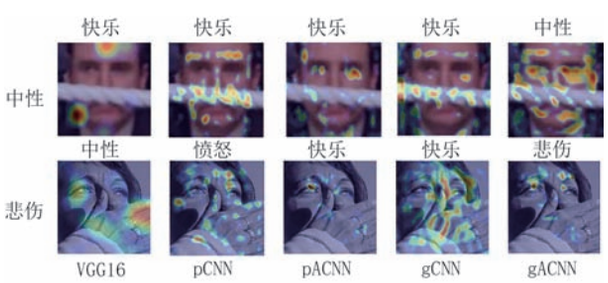


图18 使用Grad-CAM生成的具有真实遮挡的图像的注意力图[68](其中图像的表情标签在其左侧,不同方法的预测结果位于图片上方;CNN代表卷积神经网络,ACNN代表具有注意力机制的CNN[68],'p'代表基于局部图像块,'g'代表集成图像块级别的局部表示和图像级别的全局表示)

DAN,发现DAN可以关注表情相关的多个面部区域,这在提供了人类可理解的可视化解释的同时也评估了DAN相对于单头注意力模型的优势.上述算法均使用注意力机制对特征的重要性进行学习,找出那些对表情识别更为重要的特征;同时,CAM方法的使用令特征重要性能够以人类可理解的形式展现出来,这也使得研究者可以更好地解释模型.

4.3 生成通道重要性查询向量

另一类使用注意力机制提高模型可解释性的方法是使用注意力机制生成通道重要性查询向量,以描述特征图中每个通道的权重.Jiao等人[71]提出了一种基于面部注意力的卷积神经网络(Facial

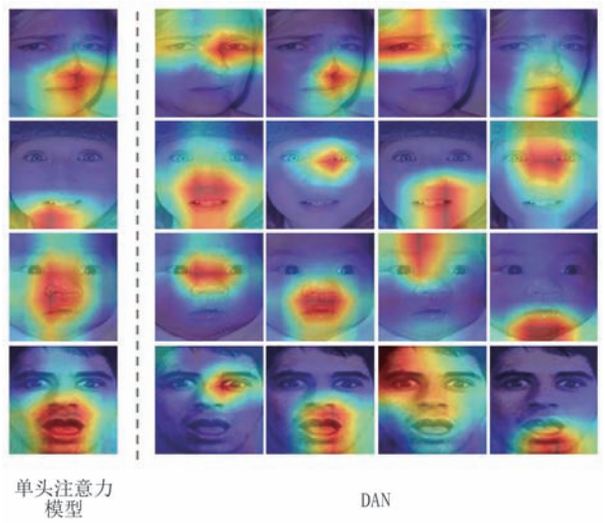


图19　使用Grad-CAM++生成的关于单头注意力模型和DAN的注意力图[69]（其中第一列根据用单头注意力模型DACL[70]生成的，其余列由DAN[69]的四个注意力头生成）

Attention based Convolutional Neural Network，FA-CNN），用于2D+3D表情识别．具体而言，模型使用VGG提取特征之后，采用全局平均池化操作，然后是用于特征重加权的编码器-解码器结构，在自动编码器结束时，将合并的特征向量转换为描述全局特征通道重要性的查询向量q，从另一个角度来看，q可以被视为卷积核，它描述了全局特征映射中每个通道的权重，最后，将原始全局特征和q卷积得到注意力图．图20展示了模型生成的区分区域的可视化，可以看出通过面部注意机制估计的有区别的面部部分具有高度的可解释性，与人类感知一致．类似地，Jiao等人[72]在2D+3D表情识别中，从信息论的角度提出了一种新的几何图生成技术，以增强从较大的受试者变化中产生的轻微的3D表情差异，并利用注意力机制，自动定位用于多尺度学习的细微判别性面部部位．具体而言，文献[72]中的注意力机制同样生成了通道重要性查询向量，并以Hadamard乘积的形式利用查询向量重新加权原始全局特征，从而指示相应通道全局特征的重要性，这一过程也增强了模型的可解释性．而Chen等人[73]提出的时空和通道注意力模块（Spatial-Temporal and Channel Attention Module，STCAM），在通道方向计算了注意力图，从而沿着通道维度增强特征；生成的通道注意力特征图为之后计算时空注意力提供了有用的信息，并最终帮助提升表情识别的性能与生成人类可理解的可视化表示．

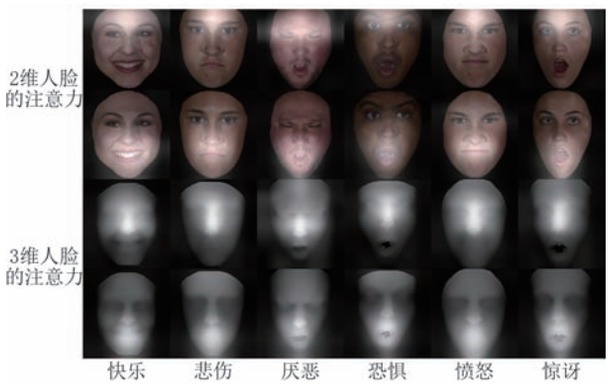


图20　FA-CNN生成的区分区域的可视化[71]

4.4　生成具特殊意义的表情图

利用注意力机制的表情识别可解释性研究还可以通过生成具有特殊意义的表情图，以辅助识别和解释模型．

例如Zhang等人[74]通过生成一种具有特殊意义的特征图，以描述特定表情类别的统一可视化．具体而言，作者提出一种包含三个模块（表情特征提取器、表情掩码精炼器和表情模式图生成器）的面部表情识别方法，算法中应用了注意力机制，并提出表情模式图的概念，它揭示了特定表情的基本特征，可以被视为衡量嵌入是否足以区分表情识别的标准，如图21所示，提供对不同面部身份不变的特定表情类别的统一可视化，提高面部表情识别的可解释性．

而Marrero Fernandez等人[75]将注意力集中在人脸上，并使用高斯空间表示进行表情识别；具体而言，模型分为两个互补的组件：第一个组件使用编码器-解码器结构的注意力模块和卷积特征提取器，并将其输出G_{att}和G_{fi}按像素相乘以获得特征注意力图，如图22所示，第二个组件负责获取面部表情的嵌入表示和分类．所生成的特征注意力图不仅有助于提升表情识别性能，还为模型提供了人类可理解的可视化解释．

Liu等人[76]则着眼于动态表情识别中的视频序列帧冗余问题，提出了一种基于片段感知的情感丰富特征学习网络（Clip-aware Emotion-richFeature Learning Network，CEFLNet），将一个视频划分为几个片段，并获得每个片段的表情强度及情感丰富表示．具体而言，CEFLNet构建了一个基于片段的特征编码器，使用级联的自注意和局部-全局关系学习模块对视频片段的时空特征进行编码，并设计了一种情绪强度激活网络来生成情绪激活图，用于定位显著的情绪片段并获得片段感知的情绪丰富表示，进而实现表情分类．情绪激活图在辅助动态表

图21 文献[74]中不同数据集中的示例图片与对应的表情模式图

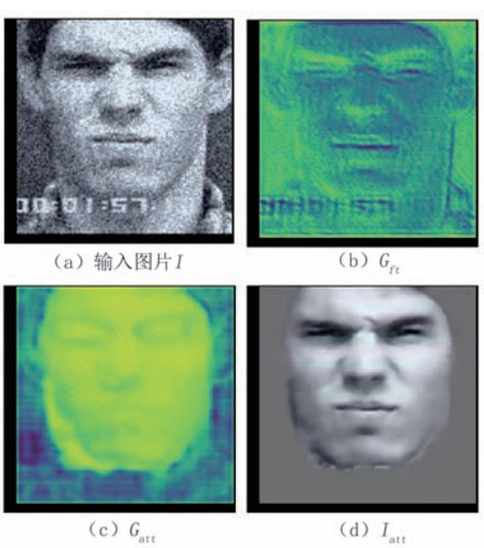


图22 文献[75]生成的注意力图I_{att}(其中(a)是噪声输入图像,(b),(c)分别是特征提取模块和注意力模块的输出,(d)是二者组合得到的注意力图I_{att})

情识别的同时也以人类理解的方式定位了情绪显著的片段,提升了模型的可解释性.

5 基于特征解耦与概念学习的可解释方法

在本节中,主要总结了特征层面表情识别可解释性方法,通过分析特征对模型的重要性和与表情的相关性,达到对模型的解释.这类方法属于机理可解释方法,主要的方法包括基于特征解耦与基于概念学习的可解释性方法.

5.1 基于特征解耦的方法

特征解耦方法希望减小各特征维度之间的相关性.特征解耦的前提是学习特征,解耦则是分离出与任务相关的特征.解耦的特征中,每一个维度都表示具体的、不相干的意义,而其中重要的一点是要让学到的特征具备人类可理解的意义.在表情识别方法中,应用特征解耦方法,可以有效地寻找人类可理解的表情特征,提高表情识别的可解释性.

本节中总结了特征解耦机制在可解释表情识别方法中的应用,包括将表情信息与其他变化因素解耦,以及将不同表情类别所特有或公有的面部区域解耦两类.

5.1.1 将表情信息与其他变化因素解耦

表情识别中常常应用特征解耦方法将表情信息与其他变化因素,如姿态、身份、年龄、光照条件等,相解耦.Ruan等人[77]提出了一种FER深度扰动分离学习方法(Deep Disturbance-disentangled Learning,DDL),能够利用多任务学习和对抗性迁移学习,同

时明确地分解多个干扰因素，如姿势、照明、身份等，DDL的训练包括干扰特征提取模型与干扰分解模型两个阶段. Halawa等人[78]使用对抗式学习来学习数据的不同变化因素的解耦表示；具体而言，论文通过使用两个不同的编码器生成两种不同的表征：第一种编码为输入图像的面部表情表示，另一种编码为其他变化因素的表示，这两个编码器使用共享编码器共享一些层，用于共享面部特征. Jiang等人[79]提出了一种身份和姿势分离的面部表情识别(Identity and Pose Disentangled Facial Expression Recognition，IPD-FER)模型，将整体面部表征视为身份、姿势和表情的组合，并将这三个分量用不同的编码器编码，以学习更具鉴别性的特征表示. 图23与24分别展示了在实验室环境与自然环境数据集下，使用身份、姿势、表情三类特征重构的图片对比(在图23中，由于实验室控制条件下采集的图片姿势均为正面人脸，因此无需考虑姿势特征)，可以看

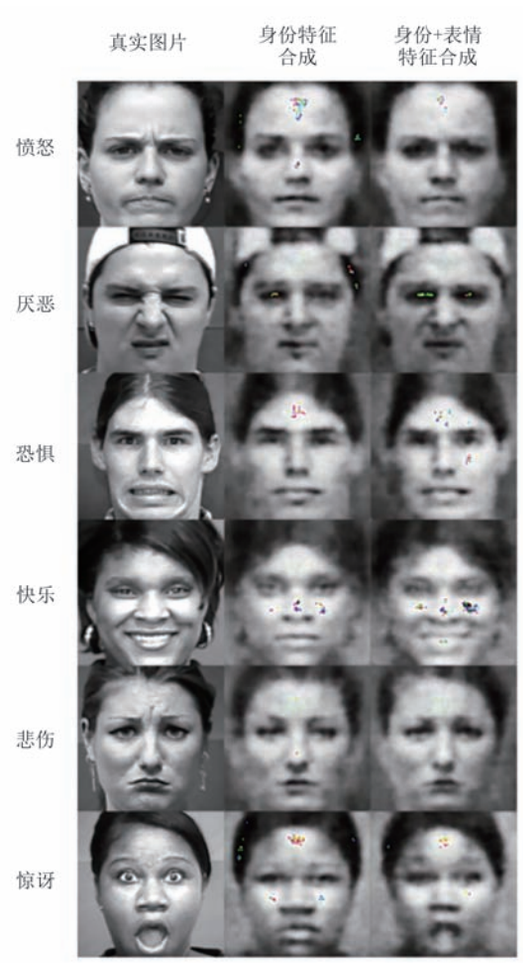


图23　文献[79]中，在实验室环境数据库CK+上对真实图像和合成图像进行比较(第一列包含真实图像，而其他列表示具有相应特征的合成图像)

出所提出的方法在解耦表情与身份、姿势特征上的有效性.

值得一提的是，许多研究尝试将表情信息与身份特征解耦. Liu等人[80]提出了身份分离的面部表情识别机(Identity-Disentangled Facial Expression Recognition Machine，IDFERM)，通过利用查询样本与其参考(例如，其挖掘或生成的正面和中性归一化人脸)的差异，从查询样本中分离身份信息的特征. 该方法包括一个硬负生成(Hard Negative Generation，HNG)网络和一个广义径向度量学习(Radial Metric Learning，RML)网络，图25展示了HNG网络生成的标准化参考面部图像，可以据此实现对身份信息的分离. Liu等人[81]从视频域中消除受试者之间的身份差异. 具体而言，作者从残差帧中推断表情，并使用预训练的人脸识别网络提取身份因子. Teng等人[82]介绍了一种典型面部表情网络(Typical Facial Expression Network，TFEN)，使用面部特征解耦器将面部特征与表情特征解耦，以最大限度地减少受试者之间面部变化的影响. 网络使用对抗性算法进行训练，通过最小化解耦后面部特征的残余影响来改进面部特征解耦器和网络性能. Li等人[83]提出的DrFER方法将解耦表示学习引入3D FER中，DrFER采用双分支框架将表情信息与身份信息解耦，并重新配置了损失函数和网络结构，以使整体框架适应于3D FER所使用的点云数据.

5.1.2　将表情类别对应的面部区域解耦

另一类方法是将特征解耦方法应用于面部区域分解，从而分解出有效区分面部表情的特征. Liu等人[86]提出了特征分解机(Feature Disentangling Machine，FDM)，用于有效地选择人脸表情特征，更重要的是，FDM将这些选定特征分解为非重叠组，特别是跨不同表情共享的公共特征和仅对目标表情有区别的表情特定特征. 具体而言，FDM将稀疏支持向量机和多任务学习集成在一个统一的框架中，并制定了新的损失函数和约束，以精确控制稀疏性并自然地分离面部表情中活跃的特征. 图26展示了用于识别CK+数据库[57]中6类基本表情的选定图像块的图示，包括了包含不同表情共享特征的图像块(绿色)和包含某个表情独有特征的图像块(蓝色或红色). 例如，对于表情对“愤怒-惊讶”，绿色框中的特征可用于识别愤怒与惊讶；而蓝色或红色框中的特征只对愤怒或只对惊讶敏感.

Xue等人[87]旨在学习与特定表情变化高度相关的，人类可解释的面部成分. 作者首先从局部深度

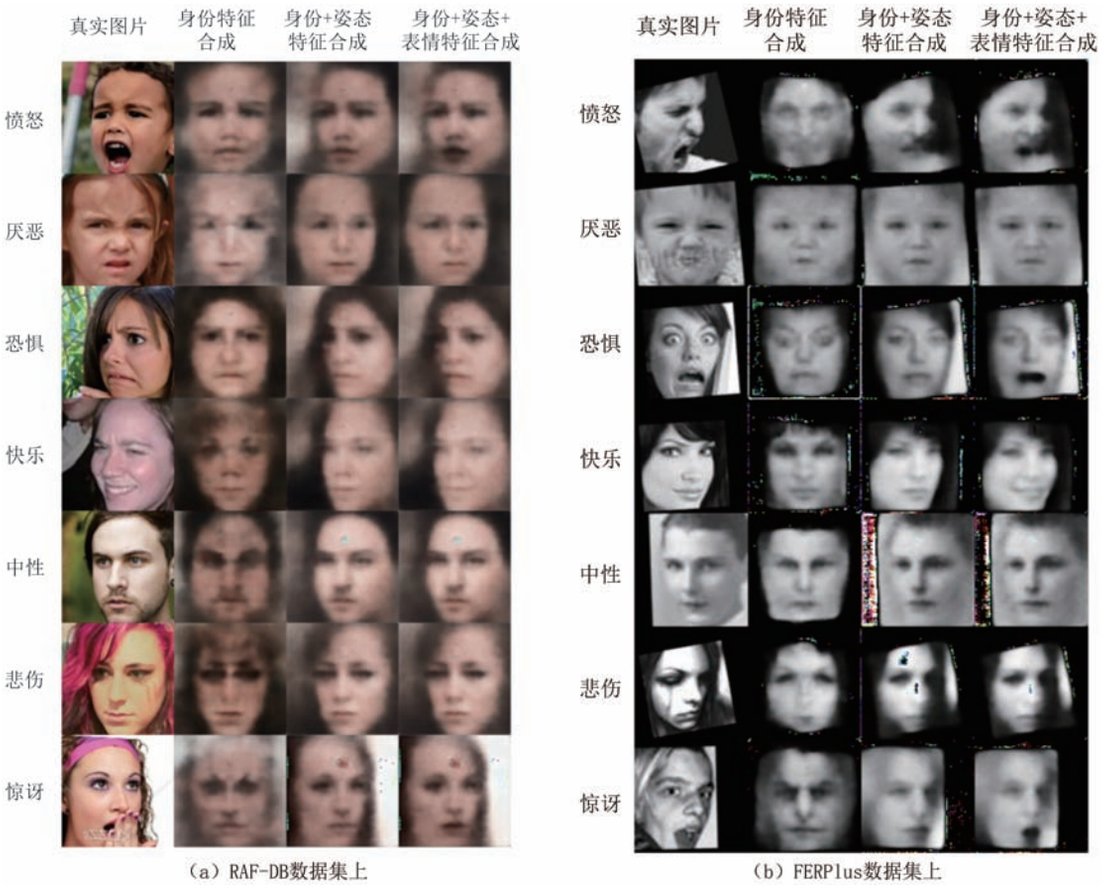


图24 文献[79]中，在自然环境数据库RAF-DB(a)与FERPlus(b)上对真实图像和合成图像进行比较(第一列包含真实图像，而其他列表示具有相应特征的合成图像)

图25 文献[80]中HNG网络的输入输出对比(输入图片来自CK+[57]、MMI[84]和Oulu CASIA[85]数据集，具有不同面部表情；输出为标准化的参考面部图像)

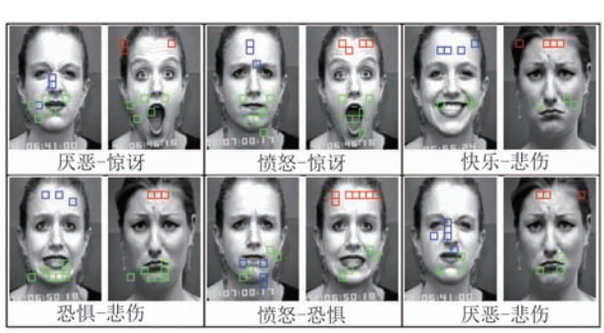


图26 文献[86]中用于识别CK+数据库中6类基本表情的选定图像块的示例

图像块序列中提取时空特征来表示面部表情动态；然后提出了一种两阶段的特征选择过程，以确定能够最好地区分表情的面部成分；最后为了验证所得到的面部成分的有效性，将来自相应区域的表情敏感特征输入到用于面部表情识别的分层分类器中以获得识别结果. 文献[87]中学习到的特征可以用人类可以理解的术语来描述，甚至可以在脸上视觉显示，具有较高的可解释性.

Li等人[88]提出了一种自监督排他性包容性交互学习(Self-supervised Exclusive-Inclusive Interactive Learning，SEIIL)方法，其中包括了一个情绪解耦结构，以产生排他性和包容性的情绪表征，通过分解

不同情绪来处理耦合的情感. 同时，作者开发了一个条件对抗性交互式学习模块，以实现两种表征之间的交互式学习，这保证了协作但有区别的表征学习.

5.2　基于概念学习的方法

基于概念学习的方法利用人类可理解的概念对模型进行解释，通过对特定概念的选择和学习，使研究者理解模型的内部特征或机理.

概念激活向量(Concept Activation Vectors，CAV)的主要思想是衡量一个概念在模型输出中的相关性. Google研究团队概述了一种名为Testing with CAV(TCAV)[89]的线性可解释方法，使用方向导数来量化用户定义的概念对分类结果的重要程度. Asokan等人[90]讨论了使用TCAV进行多模态情感识别时神经网络的可解释性问题. 为了分析模型的潜在空间，作者定义了情感AI特有的人类可理解概念，然后评估了这些概念在双向上下文长短期记忆网络(Bi-directional Contextual LSTM，BC-LSTM)多层上的影响，以表明用于情感识别的神经网络的推理过程可以使用人类可理解的概念来表示. 最后作者对提出的概念进行假设检验，以表明它们对该任务的可解释性具有重要意义. 图27展示了情感识别的TCAV方法概述.

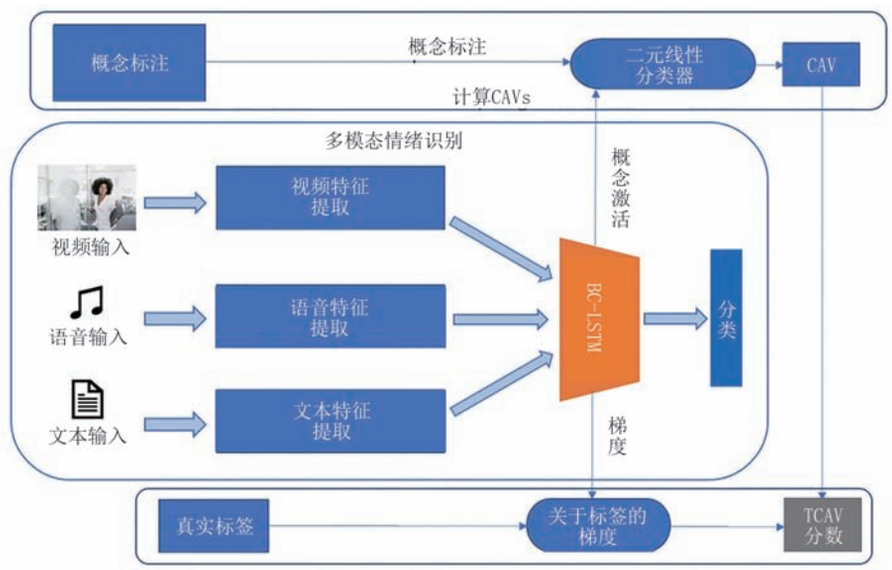


图27　情感识别的TCAV模型[90]概述

除此之外，Li等人[91]提出了一种利用语义概念识别表情的方法，作者基于公理模糊集框架，将人脸特征转化为语义概念，并利用概念区分表情. 使用概念的方法不仅降低了表情特征的维数，而且连接了图像特征和语义概念，有助于提高模型的可解释性.

6　可解释性分类方法

本节主要介绍可解释分类方法，对于表情识别中应用的可解释分类器进行了整理，主要包括两类：使用多核支持向量机的方法，以及使用决策树与深度森林的方法. 可解释性分类方法尝试对模型决策原理进行解释，属于模型可解释方法.

6.1　使用多核支持向量机

多核支持向量机(Multiple Kernel Support Vector Machine，MKSVM)[21,92]是在传统支持向量机(Support Vector Machine，SVM)的基础上，将多个核函数以线性组合的方式进行融合，然后基于该组合核函数训练一个SVM分类器. 设$K_1, K_2, \cdots, K_m$是m个核矩阵，其中$K_i=\left[k_i(x_i, x_j)\right]_{i,j=1,\cdots,n}$从不同的特征或来源获得，则MKSVM的表达式如式(1)所示. 在SVM中，数据表示是通过核函数隐式选择的，而MKSVM组合了多个核函数，并通过训练得到不同的权重，显示出不同的核函数对于分类的贡献程度，从而提高模型的可解释性.

$$K_\beta=\sum_{i=1}^{m}\beta_i K_i \tag{1}$$

在表情识别问题上应用MKSVM具有独特作用与意义. 首先，相较于一般分类问题，表情识别更为主观，并且不同类别间的差别更加细微，这使得表情识别问题相较于一般分类问题更加复杂，分类难度更高. 而多核SVM相较于单核SVM，处理复杂问题的能力更强. 同时，MKSVM通过使用不同的

核函数捕捉数据中的不同特征，也使得其能够更好地处理表情识别中类内多样性的问题．当然，MKSVM更好的可解释性可以使表情识别算法在具有较好的性能的同时具有可解释性．

Zhang等人[92]提出了一种识别疼痛表情的方法，使用MKSVM提高了决策函数的可解释性．具体来说，作者首先采用有监督的局部保留投影（Supervised Locality Preserving Projections，SLPP）提取疼痛表情特征，以解决局部保留投影算法忽略类内局部结构的问题；然后使用多核线性混合支持向量机识别疼痛表情，通过联合优化分类器的系数和核的权重，提高了决策函数的可解释性和分类器性能．类似地，Huang等人[93]使用SLPP从人脸图像中提取疲劳表情的有效特征，然后采用MKSVM对疲劳表情进行识别，与普通SVM相比，提高了决策函数的可解释性．

6.2 使用决策树与深度森林

决策树是一类经典的阐述模型可解释性的方法，而深度森林（Deep Forest，DF）则是以决策树为基本组件的一种深度学习框架．图28展示了级联深度森林的基本结构[94]，图中假设级联的每个级别由两个随机林（黑色）和两个完全随机林（蓝色）组成；并假设有三类可预测，因此每个林将输出一个3维类向量，然后将其串联以重新表示输入．以决策树为组件的DF能够捕获中间信息并跟踪决策路径，从而具有较强的可解释性．

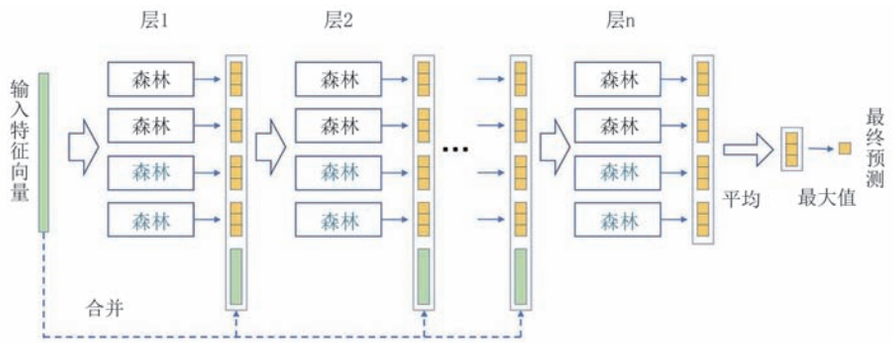


图28 DF模型[94]的整体工作流程

在表情识别领域应用DF具有一些优点．首先，正如上面所说明的，DF模型具有很好的可解释性．其次，相较于一般的分类问题（例如ImageNet数据集[95]具有约1500万张图片），表情识别的数据集数据体量一般更小，有些数据集甚至只有几百个样本．而相较于深度神经网络（Deep Neural Network，DNN），DF需要调整的参数更少，并且它的层数可以根据性能是否收敛而自适应设置，对数据集的大小更不敏感，因此DF在数据体量较小的表情识别任务上具有一定优势．

Lin等人[96]提出了一种基于纠错输出码（Error-Correcting Output Codes，ECOC）的新型DF的方法，称为EDF，用于微表情识别．如图29所示，在EDF中，DF被用作ECOC单元的基础学习器，同时，借助不同的策略，即一对一（One-vs-One，OVO）和一对全（One-vs-ALL，OVA），ECOC单元具有很大的多样性，可以从多个角度总结类别之间的差异．EDF是一种可解释的方法，如图30所示，

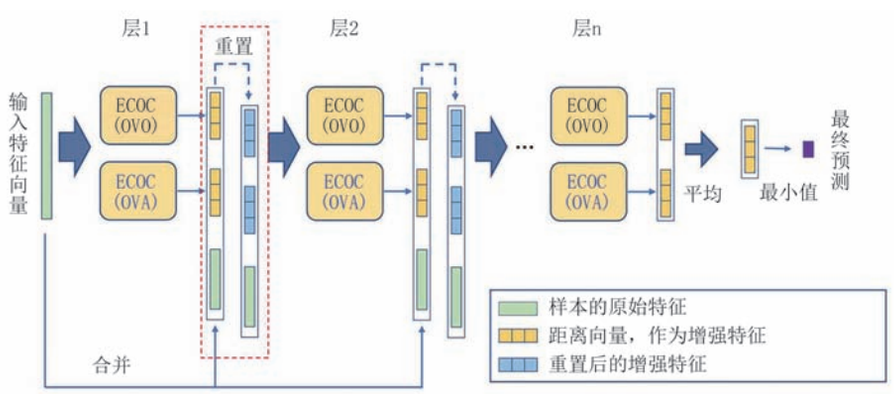


图29 EDF模型[96]的整体工作流程

借助DF，它可以通过跟踪和融合决策路径来重构特征值，以便从中轻松地确定重要特征．同时，根据反向重建过程的结果，可以总结类之间的差异．

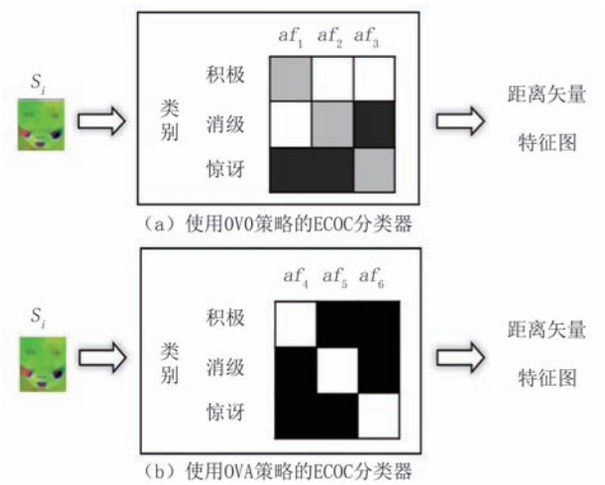


图30　文献[96]中两种类型的ECOC分类器（白色/黑色块分别表示类别分配给+1/-1组．距离矢量显示了从光流特征S_i到每个类的距离，特征图显示了哪些特征是活跃的，提供了可解释的信息．af_i表示特定二进制类任务的随机森林（例如，af_1负责处理仅涉及积极类和消极类的分类任务，而af_4负责区分消极类和其他类））

除此之外，Kim等人[97]以决策规则集的形式量化深度随机森林的特征贡献和频率，而特征贡献有助于确定特征如何影响规则集中的决策过程，因此作者借助特征贡献提出了规则消除方法，简化并解释了模型，且使模型在表情识别时可以保持较为稳定的性能．

7　实验结果比较与分析

本节中，首先对各类可解释表情识别方法进行了总结和比较，并分析了目前方法存在的一些问题．

7.1　实验结果

随着表情识别任务的不断发展，表情识别方法逐渐由使用实验室环境中的数据发展到使用更为困难、但也更具有实际意义的自然环境（如电影、互联网等）中的数据．表2、表3中分别列出了目前实验室和自然环境数据集上针对静态图像可解释表情识别方法的比较；而表4、表5则分别列出了实验室和自然环境下动态视频中可解释表情识别方法的比较，其中均包括了方法所属的可解释类型，以及在不同数据集上的准确度（括号中为其识别表情的类别

表2　实验室环境数据集上可解释静态表情识别方法的比较

数据集	方法与文献	可解释类型	准确度/%
AR face[100]	文献[47]	基于人脸基本结构	86.6(3类)
JAFFE[58]	DAM-CNN[56]	基于注意力机制	99.32(6类)
	FDM[86]	基于特征解耦	89.7(7类)
	SEIIL[88]		91.89(7类)
CK+[57]	AUDN[27]	基于人脸基本结构	92.05(8类)
	Zero-bias CNN[28]		98.3(6类)/96.4(8类)
	MSAU-Net[29]		99.1(7类)
	SG-DSN[43]		99.23(7类)
	文献[40]		97(CNN)/88(FAU+MLP)(6类)
	DAM-CNN[56]	基于注意力机制	95.88(6类)
	FMPN[62]		98.06(7类)
	gACNN[101]		96.4(7类)
	文献[74]		98.06(7类)
	FERAtt[75]		90.30(8类)
	DDL[77]	基于特征解耦	99.16(7类)
	IPD-FER[79]		99.28(6类)/98.65(7类)
	IDFERM[80]		98.35(7类)/97.76(8类)
	FDM[86]		97.7(6类)
	SEIIL[88]		98.77(7类)
	文献[91]	基于概念学习	85.60(4类)
FEI[102]	文献[91]	基于概念学习	89.25(2类)
MMI[103]	AUDN[27]	基于人脸基本结构	74.76(7类)
	MSAU-Net[29]		86.5(6类)
	SG-DSN[43]		82.64(6类)
	FMPN[62]	基于注意力机制	82.74(6类)
	DDL[77]	基于特征解耦	83.67(6类)
	IDFERM[80]		81.13(6类)
	SEIIL[88]		79.42(7类)
TFD[104]	Zero-bias CNN[28]	基于人脸基本结构	89.8(7类)
Oulu-CASIA[85]	SG-DSN[43]	基于人脸基本结构	89.24(6类)
	DDL[77]	基于特征解耦	88.26(6类)
	IDFERM[80]		88.25(6类)
痛苦表情数据集[92]	文献[92]	可解释性分类方法	90.55(2类)
BU-3DFE[105]	FA-CNN[71]	基于注意力机制	89.11(6类)
	文献[72]		89.72(6类)
	FERAtt[75]		82.11(7类)
	DrFER[83]	基于特征解耦	89.15(6类)
Bosphorus[106]	文献[72]	基于注意力机制	83.63(6类)
	DrFER[83]	基于特征解耦	86.77(6类)

表3 自然环境数据集上可解释静态表情识别方法的比较

数据集	方法与文献	可解释类型	准确度/%
300W[107]	文献[44]	基于人脸基本结构	56.63(7类)
Helen[108]	文献[44]	基于人脸基本结构	58.54(7类)
LFPW[107]	文献[44]	基于人脸基本结构	56.30(7类)
FER2013[60]	MSAU-Net[29]	基于人脸基本结构	78.3(7类)
	DAM-CNN[56]	基于注意力机制	66.2(7类)
FERPlus[109]	TransFER[63]	基于注意力机制	90.83(8类)
	EAC[65]		89.64(8类)
	IPD-FER[79]	基于特征解耦	88.42(8类)
BAUM-2i[59]	DAM-CNN[56]	基于注意力机制	67.92(6类)/61.52(7类)
RAF-DB-basic[110]	CLIPER[22]	基于文本描述	91.61(7类)
	MSAU-Net[29]	基于人脸基本结构	75.8(7类)
	SG-DSN[43]		87.13(7类)
	TransFER[63]	基于注意力机制	90.91(7类)
	EAC[65]		89.99(7类)
	gACNN[101]		85.07(7类)
	DAN[69]		89.70(7类)
	文献[74]		87.10(7类)
	DDL[77]	基于特征解耦	87.71(7类)
	IPD-FER[79]		88.89(7类)
	SEIIL[88]		88.23(7类)
RAF-DB-compound[110]	MSAU-Net[29]	基于人脸基本结构	50.2(11类)
AffectNet[111]	CLIPER[22]	基于文本描述	66.29(7类)/61.98(8类)
	MSAU-Net[29]	基于人脸基本结构	71.2(7类)
	FMPN[62]	基于注意力机制	61.52(7类)
	TransFER[63]		66.23(7类)
	EAC[65]		65.32(7类)
	gACNN[101]		58.78(7类)
	DAN[69]		65.69(7类)/62.09(8类)
	文献[74]		62.10(7类)
	文献[78]	基于特征解耦	60.53(7类)
	IPD-FER[79]		62.23(7类)
SFEW[61]	AUDN[27]	基于人脸基本结构	26.14(7类)
	MSAU-Net[29]		57.4(7类)
	SG-DSN[43]		57.42(7类)
	DAM-CNN[56]	基于注意力机制	42.30(7类)
	DAN[69]		53.18(7类)
	DDL[77]	基于特征解耦	59.86(7类)
	IPD-FER[79]		58.43(7类)

数),为对这一领域感兴趣的研究者提供参考.

下面,我们将进一步对不同的表情识别可解释性方法进行比较与分析.首先,基于文本描述和人脸基本结构的方法,在很多情况下可以获得较好的模型性能,同时能够对结果进行解释.特别是对于基于文本描述的方法,随着视觉语言模型的不断发展,其性能不断提高,并且文字模态天然可以提供人类可理解的信息,有助于提升表情识别的可解释性,因此我们认为是未来重要的发展方向之一.但是这两类模型往往需要借助其他知识或单独的模型分支从而实现可解释性.基于文本描述的方法引入文本模态以辅助表情识别与解释过程;而基于人脸结构的方法则利用人脸AU、拓扑结构等知识辅助.这使得模型需要更多样化数据,在一定程度上增加了数据获取和模型训练的难度.

注意力机制和特征解耦是另外两类较为常用的

表4　实验室环境数据集上可解释动态表情识别方法的比较

数据集	方法与文献	可解释类型	准确度/%
CK+[57]	TMSAU-Net[29]	基于人脸基本结构	99.5(7类)
	STSGN[42]		98.63(7类)
	STCAM[73]	基于注意力机制	99.08(7类)
	IFERCV[81]	基于特征解耦	97.85(7类)
	TFEN[82]		98.78(7类)
	sLMRF[97]	可解释性分类方法	92.5(6类)
MMI[103]	TMSAU-Net[29]	基于人脸基本结构	99.1(6类)
	STCAM[73]	基于注意力机制	82.21(6类)
	CEFLNet[76]		91.00(6类)
	TFEN[82]	基于特征解耦	81.73(6类)
Oulu-CASIA[85]	TMSAU-Net[29]	基于人脸基本结构	87.4(6类)
	STSGN[42]		87.23(6类)
	STCAM[73]	基于注意力机制	91.25(6类)
	TFEN[82]	基于特征解耦	91.67(6类)
BU-3DFE[105]	CEFLNet[76]	基于注意力机制	85.33(6类)
BU-4DFE[112]	文献[87]	基于特征解耦	96.64(6类,基于起始帧)/82.80(6类,基于所有帧)
CASMEII[113]+SMIC[114]+SAMM[115]	EDF[96]	可解释性分类方法	77.15(3类)

表5　自然环境数据集上可解释动态表情识别方法的比较

数据集	方法与文献	可解释类型	准确度/%
AFEW[116]	CLIPER[22]	基于文本描述	56.43(7类)
	EmoCLIP[25]		46.19(7类)
	TMSAU-Net[29]	基于人脸基本结构	47.6(7类)
	CEFLNet[76]	基于注意力机制	53.98(7类)
	IFERCV[81]	基于特征解耦	51.86(7类)
DFEW[117]	CLIPER[22]	基于文本描述	70.84(7类)
	A^{3}lign-DFER[24]		74.20(7类)
	EmoCLIP[25]		62.12(7类)
	CEFLNet[76]	基于注意力机制	65.35(7类)
	TFEN[82]	基于特征解耦	56.60(7类)
FERV39k[118]	CLIPER[22]	基于文本描述	51.34(7类)
	A^{3}lign-DFER[24]		51.77(7类)
	EmoCLIP[25]		36.18(7类)
MAFW[119]	A^{3}lign-DFER[24]	基于文本描述	53.24(11类)
	EmoCLIP[25]		41.46(11类)

可解释性表情识别方法，能在特征和机理层面解释模型. 基于注意力机制的方法，在取得较好性能的同时对特征的重要性程度进行可视化，从而增强方法的可解释性. 但目前的注意力机制仍然存在不够精确的问题，而表情中常常会存在一些比较细微的面部表现，对于这类细微表现对应的特征，注意力机制可能无法精准地捕获. 另外，正如第4节中的一些可视化结果所展现的，在应用注意力机制时仍然可能存在错误和偏差，即模型所重点关注的特征可能并不是面部表情识别中最为关键的特征. 同样，特征解耦方法也可以提升模型的可解释性，并且有助于降低过拟合，有较好的识别效果，但此类方法也有不够精确的问题(从5.1节中的一些可视化图像中也可看出)，这可能导致信息的丢失. 另外，在评估特征解耦的效果时，目前在表情识别领域的方法大都使用识别准确度的变化来评估，而如果能加入其他一些评价指标，如灵敏度[98]、稀疏性[99]，则可以更加客观地评价方法的性能与解释性. 而对于基于概念学习的方法，在可解释性的表情识别研究中较为少见，可能是因为选择合适的概念是相对比较难的，并且可能带有主观偏见，另外概念学习也会带来计算成本.

至于应用可解释的分类方法的表情识别算法，在近年同样使用得较少，我们猜测这可能是由于为了使得模型的决策和分类过程可解释，往往需要更简单的模型结构，如多核支持向量机、决策树等，但这也在一定程度上限制了模型的性能. 因此，如何平衡模型的性能与可解释性，也是目前需要表情识别可解释性研究解决的问题之一，在下一节中我们将进一步论述这一问题.

值得注意的是，目前在表情识别领域，关于可解释性的探究仍然多为定性分析，并且表现形式较为多样，而缺少定量的、统一的评价指标，关于这一问题，将在7.1.2节中进行更为详细的论述.

7.2　问题分析

尽管目前表情识别问题的可解释性已经引起部

分学者的兴趣和讨论，但目前关于表情识别可解释性研究的相关文献仍然相对较少，算法仍然较为有限，并且具有一定的局限性.

首先，对于表情识别可解释性问题的研究仍然缺少有效的评估方式. 目前的表情识别可解释性的算法大部分仍然是定性地分析模型可解释性，而缺少定量地对模型可解释性进行度量，并且由于解释范围、解释方法原理不同等因素，缺乏统一有效的评估方式. 参考深度学习可解释性的评估方式，对于表情识别问题的可解释性评估，可以采用以下指标：

第一，失真度(Infidelity)或保真度(Fidelity)，它是用来判断解释方法选取的是否是对模型分类影响较大的输入特征，换言之，当这些特征被删除时，是否会大幅改变预测结果. 在表情识别任务中，失真度可用于评估可解释性研究是否能够找到那些与表情识别模型输出相关性较大的面部特征. 失真度的定义来源于Ancona等人[120]定义的完整性公理. 具体而言，定义模型输入 $x\in R^d$，模型输出 $f(x)$，解释函数 $\Phi(f,x)$，扰动 $I\in R^d$ 是概率分布为 μ_I 的随机变量，则失真度[98]为

$$INFD(\Phi,f,x)=E_{I\sim\mu_I}\left[(I^T\Phi(f,x)-(f(x)-f(x-I)))^2\right] \quad (2)$$

在表情识别问题中，所添加的扰动可能是更改人脸图像中的某些像素值，遮挡面部中的某些区域等.

第二，灵敏度(Sensitivity)，它指的是可解释性方法对于输入的波动的敏感程度. 一般情况下，我们不希望输入的微小波动导致解释的变化，因为这意味着解释不够鲁棒甚至不够可靠. 特别是在表情识别问题中，因为表情本身是多变的，因此我们不希望在输入产生微小变化时解释发生大的改变. 除此之外，灵敏度可以用来评估噪声对解释方法的影响，而噪声在表情数据，尤其是真实世界表情数据中是广泛存在的，另外对抗性攻击也会引入对抗性噪声. Yeh等人[98]对解释的最大灵敏度的定义如式(3)所示，其中 x 为输入，$f(x)$ 为模型输出，$\Phi(f,x)$ 为可解释性方法函数，r 为给定的输入区间半径. 同时，他们使灵敏度可以和失真度结合使用，以避免解释对输入的变化过于不敏感的情况.

$$SENS_{MAX}(\Phi,f,x,r)=\max_{\|y-x\|\leqslant r}\|\Phi(f,y)-\Phi(f,x)\| \quad (3)$$

第三，稀疏性(Sparsity)，即模型应该捕捉那些对分类重要的特征，而尽量排除不相关的特征. 在面部数据中，往往包含着多种信息，身份、姿态、光照条件等等都可能是表情识别的干扰因素. 换言之，表情识别问题中常常存在着大量的干扰信息. 表情识别可解释性研究应该尽可能忽略不相关的特征，而找到对识别具有重要作用的特征，当然，这些特征应该尽可能是稀疏的. 具有稀疏性的可解释方法更加简洁，使人更易于理解，并且排除了干扰信息，同时节约了计算成本. Yuan等人[99]给出的关于稀疏性的定义如式(4)所示，其中 N 表示图片数，$|m_i|$ 表示选取的特征数，$|M_i|$ 表示特征总数.

$$Sparsity=\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{|m_i|}{|M_i|}\right) \quad (4)$$

其次，表情识别模型需要平衡准确性和可解释性. 在一般规律中，模型的复杂度和模型的准确度在一定程度上呈正相关，然而越复杂的模型，往往越难以对其进行解释[121-122]，例如决策树等简单的可解释模型在复杂的场景和模型中往往很难达到很好的效果. 尤其表情识别任务往往比一般分类任务具有更高的类间相似性和类内差异[110,123-124]，这也意味着其具有较高的分类难度，因此，平衡模型准确度和可解释性是目前研究需要解决的问题之一. 一方面，可以通过简化模型的方式(如模型剪枝[125-126])，在确保性能可接受的前提下提升模型的可解释性. 另一方面，模型的可解释性和准确性并不是完全对立的关系，对模型合理的解释，可以在一定程度上避免模型错误的学习与发展方向，从而降低模型出现错误结论的可能性，或发现并解决黑盒模型潜在的严重错误[127]. 在表情识别问题中，可以利用可解释性研究使模型更多地关注真正与表情相关的面部区域中，而减轻身份、姿态等干扰因素的影响.

同时，表情识别可解释性具有数据有限的问题. 近年来，随着表情识别任务吸引越来越多的研究者的关注，表情识别的数据集也在逐步发展和完善，有EmotioNet[37]、RAF-DB[110]、AffectNet[128]等多个较为大型的数据集. 然而，相较于其他任务，表情识别的数据和数据集仍然较为有限，而具有可解释的额外标注的表情识别数据集则更少. 因此，训练出性能较好的模型解释器，尤其是以有监督的方法训练解释器，由于缺少大量的样本与标注，仍然较为困难. 一种较为可行的方法是使用半监督或无监督的学习方法，利用已有的表情识别数据集与数据标注，结合大量的无标注数据，对于表情识别可解释问

题进行研究.

8　未来展望

本节将在对目前表情识别问题以及可解释性研究方法的分析的基础上，对未来这一领域可能的发展方向进行展望，为感兴趣的研究者提供参考.

8.1　复杂表情识别的可解释性

目前的可解释表情识别方法主要针对6类基本表情，即惊讶、恐惧、厌恶、快乐、悲伤和愤怒，或加入中性类的7类基本表情. 然而，人类表情往往更加复杂和多样. 为此，Du等人[129]提出了复合表情的概念，指一些由多种基本表情组合而成的复杂表情. 其后研究者们针对复合表情识别问题进行了不断探索. Li等人[110]提出了真实世界情感面部数据库RAF-DB，其中包含了组合了两类基本表情的复合表情，如图31所示，并提出一种方法，通过在最大化类间散射的同时保持局部接近性来增强深度特征的判别能力，可以用于基本和复合表情识别. Liu等人[130]提出了一种称为“Boosting POOF”的自动框架，使用从局部面部区域提取的低级特征（如LBP、HOG、SIFT和Gabor）来提取差异化的中级特征，该框架适用于基本和复合表情识别. Liu等人[119]提出了一个包含10，045个自然环境视频音频片段的大型多模态复合情感数据库MAFW，并介绍了一种基于Transformer的表情片段特征学习方法，利用不同情绪和模态之间的表情变化关系来识别复合情绪. Li等人[131]提出了多标签面部表情数据库RAF-ML，并使用深度双流形卷积神经网络，通过共同保持深度特征的局部亲和性和标签的流形结构来学习多标签表情的判别特征.

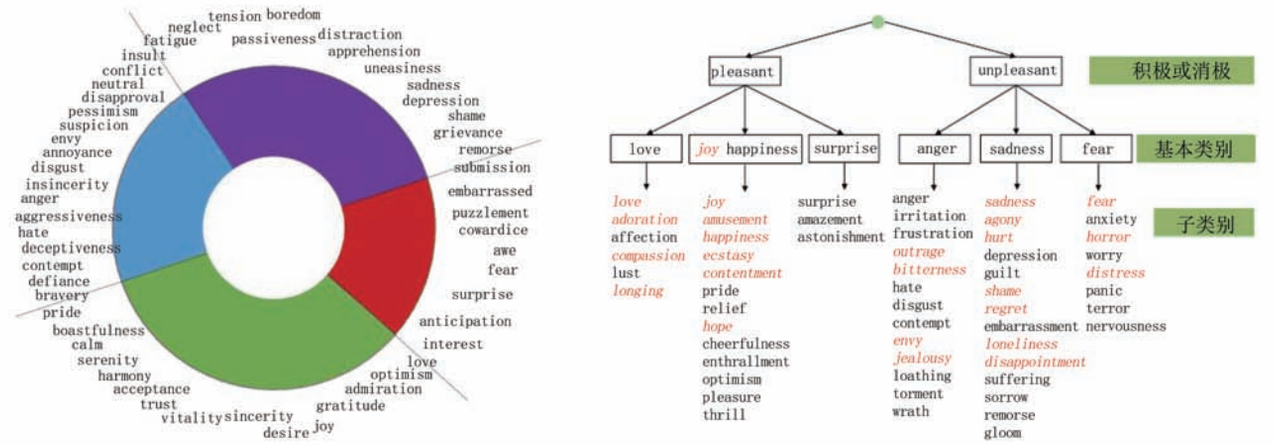
图31　数据集RAF-DB[110]中的复合表情示例

复合表情可以看作是6类基本表情的延伸，并不足以完全阐述不同情绪下面部表情之间的微妙之处. 因此，研究者提出细粒度表情识别. 细粒度表情一般依托于心理学领域的情感理论如“情感之轮”理论[132]和情感层次模型[133]，旨在识别大量更加细致微妙的表情类别，如尴尬、紧张、骄傲等. Wang等人[134]创建了一个包含54种细粒度表情的数据集F^2ED，如图32(a)所示，并提出了一种面部姿势生成对抗网络来合成新的面部表情图像以增强用于训练的数据集，然后学习了用于表情分类的模型Fa-Net. 不同于F^2ED在受控环境中收集数据，Liang等人[29]提出了自然环境下的细粒度表情基准FG-Emotions，如图32(b)，包含33个表情类别的图像和视频，并使用一种基于端到端多尺度动作单元的网络，用于图像面部表情识别，然后将其进一步扩展为双流模型用于视频表情识别.

复杂表情识别较之基本表情识别难度更大，模型结构也往往更为复杂，因此目前针对可解释复杂表情识别的研究[22,29,37,135]仍然较少，但随着可解释性研究的不断发展，对于复杂表情识别可解释性的研究将是非常值得研究的课题.

8.2　多模态情绪识别的可解释性

人脸表情并非情感的唯一表达方式，语音、文字、人体生理信息等也可以用于辅助情绪识别. 通过与其他模态信息进行互补，有助于提高情绪识别模型的准确度与鲁棒性. 因为人脸表情是很微妙的，并且在某种情况下可能会存在误导性，人们在某些情况下可能会隐藏真实表情，甚至做出虚假的表情，如所谓的“强颜欢笑”，而这时如果加上语音或者生理信息的辅助，则更容易对受试者的情绪做出更准确的判断. 随着多模态情绪识别模型的不断发展，对于多模态情绪识别可解释性的研究也逐渐吸引了研究者的兴趣. 多模态情绪识别可解释性的研究也有助于研究者理解模型的判断依据和内在逻辑，使得模型结果更加可信，并且多模态模型的可解释性研究使研究者在了解模型的基础上也更容易发



(a) F²ED使用的轮状结构及其类别标签　　(b) FG-Emotions使用的层次结构及其类别标签

图32　数据集F²ED[134]及FG-Emotions[29]中的细粒度表情类别. 其中(a)是F²ED所使用的轮状结构,以及据此可分为4大类的54种细粒度表情. (b)是FG-Emotions所使用的层次结构,其中的叶子结点代表细粒度表情. 注意作者根据细粒度分类任务的实际需求,挑选了33类表情(图中为黑色字表示的叶子结点)作为数据集的标签

掘和利用模态间的互补性质. 已经有一些研究对多模态情绪识别的可解释性进行了探索. Zadeh等人[136]将构建n模态动态定义为一个分层过程,并提出了一种名为动态融合图(Dynamic Fusion Graph, DFG)的新融合模型. 通过图连接效率,DFG很容易解释. Tsai等人[137]提出了多模态路由(Multimodal Routing),它针对每个输入样本以不同的方式动态调整输入模态和输出表示之间的权重,因此可以识别单个模态和跨模态特征的相对重要性. 通过路由进行权重分配使作者不仅能够在数据集的整体趋势上解释模态预测关系,而且还可以在局部解释每个单个输入样本的模态预测关系. Asokan等人[90]使用概念激活向量,定义情感人工智能特有的人类可理解的概念,以解决多模态模型的可解释性问题. Kumar等人[138]则结合了分治方法来计算表示每个语音和图像特征重要性的形状值,从而实现多模态情感系统的可解释性. Palash等人[139]提出了可解释的情境知识多模态情绪识别,用于基于表情、姿势和步态的情绪识别. 作者利用从位置类型中提取的情境知识和场景中得出的形容词-名词对,以及情绪的时空平均分布来生成解释.

与单模态的表情识别可解释性研究有所不同的是,多模态的可解释性研究关注的方向不仅包括不同模态的特征提取过程,还包括了多模态特征融合过程的解释,这也为研究者提供了新的思路. 当然,多模态模型相较于单模态模型,往往更为复杂,也为可解释性研究带来了挑战.

8.3　大模型表情与情绪识别的可解释性

近年来,大模型取得了快速的发展. 大模型的通用性使其在微调或者零样本的情况下可以适用于多种任务,这也为表情和情绪识别提供了更多的选择和可能性. 为了提高大模型的安全性与可信性,使其可以获得更广泛的应用,关于大模型的可解释性研究具有重要的研究价值.

8.3.1　视觉大模型的可解释性

尽管在大模型发展之初研究者主要聚焦于大语言模型,但随着大模型的不断发展,一些研究者开始关注视觉这一同样重要的模态上的大模型. Kirillov等人[140]提出了分割一切模型(Segment Anything Model, SAM),设计了一个可提示的分割任务作为预训练,得到的模型可以在新的图像分布和任务中进行零样本迁移. Bai等人[141]利用序列建模的方法学习大视觉模型,这是一个纯视觉的模型,通过定义通用的格式"视觉句子"来代表不同的注释,可处理多样的任务. 这些模型虽然没有直接应用于表情识别中,但其通用性为其迁移到情感计算领域提供了基础.

而关于大视觉模型的可解释性,Bai等人[141]提到的"类比提示"方法测试了模型的解释能力. 另外,Fu等人[142]在LVM中加入人机交互模块,利用人类知识提高模型的可靠性与可解释性. 但总体而言,LVM的可解释性研究仍具有挑战,仅依靠视觉单一模态完成对模型规模较大的LVM的解释有一定难度,尤其表情本身也较为微妙和多变,因此解释也更具挑战,而人机交互又会带来额外的人工成本. 因此,利用其他模态,如文字,是提升模型可解

释性可行的一个思路.

8.3.2　视觉语言模型的可解释性

目前的多模态模型中，视觉语言模型(Vision Language Model，VLM)[143]获得了非常多的关注与探索. 研究者通过对视觉和文字这两类广泛使用的模态的融合，获得更丰富的语义理解和更强的交互性. 本节中，我们将针对VLM的可解释性进行分析与展望.

由于自然语言对人类而言是可以直观理解的，因此VLM中语言模态的信息，可以有效地为表情识别机理提供解释. 例如Li等人[22]提出了基于CLIP框架的CLIPER，通过学习具有可解释性的文本描述符以解释表情识别模型. 类似地，Tao等人[24]设计了可学习的多维对齐标记替换CLIP中的输入标签文本，并发现其成功学习了与表情相关的单词，这也体现了其可解释性. Foteinopoulou等人[25]利用关于上下文、表情或情感线索的文本描述作为自然语言监督，有利于研究者了解模型基于哪些因素来构建决策.

目前的VLM中，有更多基于自然语言进行解释的方法. 如Lu等人[144]尝试使模型利用不同模式的可用信息来综合一致和完整的思维链(Chain of Thought，CoT)，如图33所示，作者提出了标注详细解释的多模态科学问答数据集ScienceQA，并训练模型在生成答案的同时，生成相应的推理解释，提高了其可解释性. Yang等人[145]提出了一种内心独白多模态优化方法，通过模拟内心独白过程来解决复杂的视觉语言问题，有助于提升模型的推理和解释能力. 具体而言，作者使大语言模型和视觉语言模型能够通过自然语言对话进行交互，并使用两阶段训练过程来学习如何进行内心独白(自问问题和回答问题). Yan等人[146]通过学习简洁的描述性属性进行视觉识别任务，作者提出了一种学习搜索方法来寻找简洁的属性集，然后通过这些属性来帮助视觉分类，同时这些描述性属性的使用也提高了模型的可解释性.

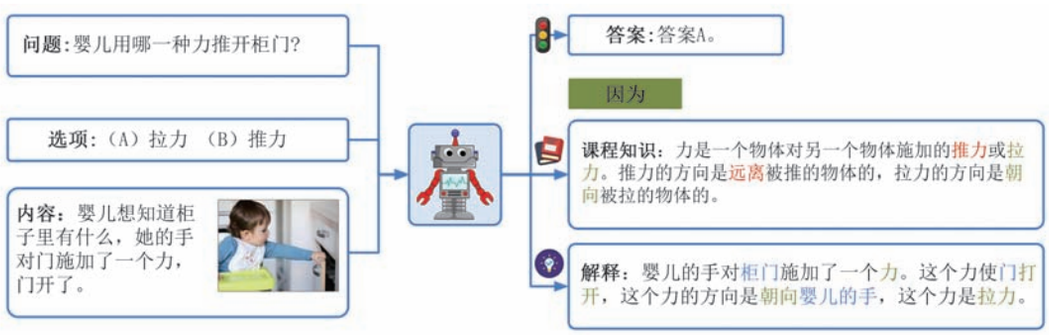


图33　文献[144]中作者使模型在预测答案的同时给出人类可理解的解释(在图中的例子中，作者希望模型对于一道选择题的回复包括了答案、题目对应的背景知识以及答案解释)

当然，在VLM的可解释性研究中，也有研究者尝试使用其他的可解释方法. 例如Yellinek等人[147]使用树结构提高VLM的可解释性；Lin等人[148]利用稳定扩散为VLM生成参考演示，从而提高视觉可解释性. 这些VLMs上可解释性的探索，也为使用VLM的情绪识别的可解释性研究提供了思路.

近年来，VLM取得了显著的进展，然而，研究表明[143,149-150]，这类模型在可信问题上仍然存在一定的漏洞，仍然存在偏见、幻觉、有毒内容等问题. 通过对其可解释性的探索，有助于研究者在了解模型决策原理的基础上解决模型的可信问题. 尤其对于情绪识别这类可能会用于医疗、安全等重要领域的任务，具有良好的可解释性的模型可以增强用户对模型的信任，也可以在一定程度上减轻模型由于虚假的相关性而导致的错误.

8.3.3　多模态大模型的可解释性

由于多模态情绪信息可以相互补充，并且多模态更加符合现实世界的实际情况，近年来越来越多的研究者针对多模态大模型展开了研究. 除了上面所提到的视觉语言模型，还有一些多模态大模型结合了更多的模态信息，例如语音、图像、视频、动作、文字等多个模态的融合.

目前许多多模态大模型依托于大语言模型(Large Language Model，LLM)，并结合处理其他模态信息的模块. 例如X-LLM[151]模型使用了“X2L”接口将多种模态(包括图像、语音和视频)的信息注入LLM中，从而使模型可以处理多模态输入. Video-LLaMA[152]分别使用视觉语言分支和音频语

言分支，将视频中的视觉与听觉信息转换为与LLM文本输入兼容的查询表示．除此之外，还有BuboGPT[153]、X-InstructBLIP[154]、CoDi-2[155]等多个模型被提出．

最近，有研究者关注于原生多模态的探索．原生多模态在设计之初即支持多模态信息，模型直接基于图片、视频、语音、文本等多模态信息进行预训练，并且一般由同一个模型端到端地完成所有任务．原生多模态对于多模态信息能够更有效地进行融合，从而在跨模态任务上展现出较好的效果．近期关于原生多模态的工作包括Gemini[156]、Unified-IO 2[157]、GPT-4o①等．

以上这些多模态大模型可以处理视觉、语音、文本等多模态数据，这也为多模态的情绪识别提供了更多的选择，并且更有利于研究者充分发挥不同模态情绪信息间的互补作用．

至于多模态大模型情绪识别的可解释性，已经有研究者进行了探索．例如Lian等人[158]提出新的任务“可解释的多模态情感推理”，根据情绪预测背后的推理过程为多模态大模型的预测提供解释，并提出一个数据集作为基准．另外，其他多模态大模型的可解释性研究，例如Yuan等人[159]基于检索增强上下文学习的方法，也提供了有益的思路．当然，由于多模态大模型参数量的增加和模型结构的复杂化，使得多模态大模型情绪识别任务的可解释性具有一定的挑战性，这也是未来研究者值得探究的方向．另外，目前多模态大模型的可解释性研究，主要是结果可解释性方法，如何利用机理可解释性和决策可解释性方法也是未来可以关注的方向．

8.4 基于可解释性提升泛化能力

基于可解释性分析提升表情识别模型的泛化能力也是一个值得探索的方向．当机器学习模型学习的是“相关性”而非“因果性”，在遇到新的数据或受到其他因素影响时往往很容易做出错误判断[160]，即模型缺乏较好的泛化性能．对于模型可解释性的研究，有助于说明模型输入到输出之间的因果关系，加深研究者对于模型本质的认识．从这个角度考虑，可解释性的研究增强了模型对于“因果性”的学习，有利于提高模型的泛化性能．

已经有研究者对可解释模型的泛化性能进行了研究．Zhou等人[42]提出时空语义图网络，通过从面部拓扑结构中进行端到端的特征学习来自动学习空间和时间模式，较以往方法具有更强的解释性和泛化能力．Liu等人[44]提出了一种基于人类视觉认知FER系统的可解释GNN模型，其中人类视觉认知策略包括“局部视觉认知”和“区域协同识别”两个关键过程，将不同区域的特征联系起来，找出原本独立部分之间的内涵关系，使得模型有更强的泛化能力．Zhang等人[74]利用可解释方法更准确地寻找表情相关区域，通过约束减少面部身份信息对表情识别的负面影响，使模型具有更好的通用性和泛化性，并通过跨数据集实验验证了其泛化性能，如表6所示，作者展示了所提出方法与gACNN[68]和SPWFA-SE[161]方法在跨数据集性能上的比较，并获得了满意的结果．

表6 文献[74]中方法的跨数据集性能比较

方法	训练数据集	测试数据集	准确度/%
gACNN[68]	RAF-DB	CK+	81.07
gACNN[68]	RAF-DB	Oulu-CASIA	50.31
gACNN[68]	AffectNet	CK+	91.64
gACNN[68]	AffectNet	Oulu-CASIA	58.18
SPWFA-SE[161]	RAF-DB	CK+	81.72
SPWFA-SE[161]	AffectNet	CK+	85.44
文献[74]	RAF-DB	CK+	83.12
文献[74]	RAF-DB	Oulu-CASIA	53.79
文献[74]	AffectNet	CK+	92.97
文献[74]	AffectNet	Oulu-CASIA	60.36

然而，到目前为止，这方面的工作仍然相对有限．因此，利用可解释性学习来提高表情识别模型的泛化能力是一个具有较大潜力和研究空间的方向．

9 总 结

本文首先简要概述了表情识别任务与可解释性研究的背景知识，然后根据结果、机理、模型可解释性的概念对可解释方法进行整理，并依据表情识别自前至后的顺序，对表情识别可解释性方法进行了整理．具体而言，文中分别从基于文本描述的可解释方法、基于人脸基本结构的表情识别方法、表情识别中的注意力机制、基于特征解耦与概念学习的可解释方法，以及可解释性分类方法几个方面，总结了可解释的表情识别方法．最后，论文对于表情识别可解释性研究进行了对比与分析，并预测了未来这一领域可能的发展方向，包括复杂表情的可

① OpenAI：Hello GPT-4o，https://openai.com/index/hello-gpt-4o/ 2024,5,13

解释性，多模态情绪识别的可解释性、大模型表情与情绪识别的可解释性以及基于可解释性提升泛化能力.

致 谢 本文受国家自然科学基金面上项目资助(批准号：62076034). 感谢对本文给出宝贵建议和支持的各位同行，感谢编辑与各位审稿人的严谨专业的评审与意见！

参 考 文 献

[1] Zhao Li-Zhuang, Gao Wen, Chen Xi-Lin. Eigenface dimension variant classification and it's application in expression recognition. Chinese Journal of Computers, 1999, 22(6): 627-632 (in Chinese)
(赵力庄，高文，陈熙霖. Eigenface 的变维分类方法及其在表情识别中的应用. 计算机学报，1999，22(6)：627-632)

[2] Jung H, Lee S, Yim J, et al. Joint fine-tuning in deep neural networks for facial expression recognition//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 2983-2991

[3] Yao A, Cai D, Hu P, et al. Holonet: Towards robust emotion recognition in the wild//Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI). Tokyo, Japan, 2016: 472-478

[4] Hou B, Zhou Z. Learning with interpretable structure from gated rnn. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(7): 2267-2279

[5] Kim B, Khanna R, Koyejo O. Examples are not enough, learn to criticize! Criticism for interpretability//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2288-2296

[6] Adler P, Falk C, Friedler S A, et al. Auditing black-box models for indirect influence. Knowledge and Information Systems, 2018, 54(1): 95-122

[7] Karpathy A, Li F-F. Deep visual-semantic alignments for generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 664-676

[8] Lipton Z C. The mythos of model interpretability. Communications of the ACM, 2018, 61(10): 36-43

[9] Zhao B, Wu X, Feng J, et al. Diversified visual attention networks for fine-grained object classification. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256

[10] Zeng Yi-Fu, Lan Tian, Wu Zu-Feng, et al. Bi-Memory based attention model for aspect level sentiment classification. Chinese Journal of Computers, 2019, 42(8): 1845-1857 (in Chinese)
(曾义夫，蓝天，吴祖峰，等. 基于双记忆注意力的方面级别情感分类模型. 计算机学报，2019，42(8)：1845-1857)

[11] Feng Z, Xu C, Tao D. Self-supervised representation learning by rotation feature decoupling/Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 10356-10366

[12] Zhou F, Hang R, Liu Q. Class-guided feature decoupling network for airborne image segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(3): 2245-2255

[13] Yang Li-Ji, Wang Jia-Qi, Jing Li-Ping, et al. Semantic representation learning of convolutional neural network based on tensor computation. Chinese Journal of Computers, 2023, 46(3): 568-578 (in Chinese)
(杨礼吉，王家祺，景丽萍，等. 基于张量计算的卷积神经网络语义表示学习. 计算机学报，2023，46(3)：568-578)

[14] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019, 267: 1-38

[15] Krishnan R, Sivakumar G, Bhattacharya P. Extracting decision trees from trained neural networks. Pattern Recognition, 1999, 32(12): 1999-2009

[16] Yang C, Rangarajan A, Ranka S. Global model interpretation via recursive partitioning//Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). Exeter, UK, 2018: 1563-1570

[17] Fan F-L, Xiong J, Li M, et al. On interpretability of artificial neural networks: A survey. IEEE Transactions on Radiation and Plasma Medical Sciences, 2021, 5(6): 741-760

[18] Setiono R, Liu H. Understanding neural networks via rule extraction//Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1. Montreal, Canada, 1995: 480-485

[19] Saad E W, Wunsch D C. Neural network explanation using inversion. Neural Networks, 2007, 20(1): 78-93

[20] Friesen E, Ekman P. Facial action coding system: A technique for the measurement of facial movement. Palo Alto, 1978, 3(2): 5

[21] Bach F R, Lanckriet G R G, Jordan M I. Multiple kernel learning, conic duality, and the SMO algorithm//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, 2004: 6

[22] Li H, Niu H, Zhu Z, et al. CLIPER: A unified vision-language framework for in-the-wild facial expression recognition. arXiv preprint arXiv:2303.00193, 2023

[23] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning (ICML). Virtual, 2021: 8748-8763

[24] Tao Z, Wang Y, Lin J, et al. A^{3}lign-dfer: Pioneering comprehensive dynamic affective alignment for dynamic facial expression recognition with clip. arXiv preprint arXiv: 2403.04294, 2024

[25] Foteinopoulou N M, Patras I. Emoclip: A vision-language method for zero-shot video facial expression recognition. arXiv preprint arXiv:2310.16640, 2023

[26] Simon T, Nguyen M H, Torre F D L, et al. Action unit detection with segment-based svms//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and

Pattern Recognition, San Francisco, USA, 2010: 2737-2744

[27] Liu M, Li S, Shan S, et al. Au-aware deep networks for facial expression recognition//Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Shanghai, China, 2013: 1-6

[28] Khorrami P, Paine T L, Huang T S. Do deep neural networks learn facial action units when doing expression recognition? // Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). Santiago, Chile, 2015: 19-27

[29] Liang L, Lang C, Li Y, et al. Fine-grained facial expression recognition in the wild. IEEE Transactions on Information Forensics and Security, 2021, 16: 482-494

[30] Tang Y, Zeng W, Zhao D, et al. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 12879-12888

[31] Shingjergi K, Iren Y, Klemke R, et al. Interpretable explainability for face expression recognition//Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence. Amsterdam, The Netherlands,2023: 10

[32] Zhao K, Chu W, De la Torre F, et al. Joint patch and multi-label learning for facial action unit and holistic expression recognition. IEEE Transactions on Image Processing, 2016, 25 (8): 3931-3946

[33] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017). LongBeach, USA, 2017: 5999-6009

[34] Tallec G, Yvinec E, Dapogny A, et al. Multi-label transformer for action unit detection. arXiv preprint arXiv:2203.12531, 2022

[35] Wang L, Qi J, Cheng J. Action unit detection by exploiting spatial-temporal and label-wise attention with transformer// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, USA, 2022: 2469-2474

[36] Li Y, Mavadati S M, Mahoor M H, et al. Measuring the intensity of spontaneous facial action units with dynamic bayesian network. Pattern Recognition, 2015, 48(11): 3417-3427

[37] Benitez-Quiroz C F, Srinivasan R, Martinez A M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 5562-5570

[38] Walecki R, Rudovic O, Pavlovic V, et al. Copula ordinal regression for joint estimation of facial action unit intensity// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 4902-4910

[39] Zhang Y, Jiang H, Wu B, et al. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Republic of Korea, 2019: 733-742

[40] Deramgozin M, Jovanovic S, Rabah H, et al. A hybrid explainable AI framework applied to global and local facial expression recognition//Proceedings of the 2021 IEEE International Conference on Imaging Systems and Techniques (IST). Kaohsiung, China, 2021: 1-5

[41] Ribeiro M T, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1135-1144

[42] Zhou J, Zhang X, Liu Y, et al. Facial expression recognition using spatial-temporal semantic graph network//Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates, 2020: 1961-1965

[43] Liu Y, Zhang X, Zhou J, et al. SG-DSN: A semantic graph-based dual-stream network for facial expression recognition. Neurocomputing, 2021, 462: 320-330

[44] Liu S, Huang S, Fu W, et al. A descriptive human visual cognitive strategy using graph neural network for facial expression recognition. International Journal of Machine Learning and Cybernetics, 2024, 15(1): 19-35

[45] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80

[46] Perkins D N. A definition of caricature and caricature and recognition. Studies in The Anthropology of Visual Communication, 1975, 2: 1-24

[47] Gao Y, Leung M K H, Hui S C, et al. Facial expression recognition from line-based caricatures. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2003, 33(3): 407-412

[48] Calder A J, Young A W, Rowland D, et al. Computer-enhanced emotion in facial expressions. Proceedings of the Royal Society B: Biological Sciences, 1997, 264 (1383): 919-25

[49] Calder A J, Rowland D, Young A W, et al. Caricaturing facial expressions. Cognition, 2000, 76(2): 105-146

[50] Leppänen J M, Kauppinen P, Peltola M J, et al. Differential electrocortical responses to increasing intensities of fearful and happy emotional expressions. Brain Research, 2007, 1166: 103-109

[51] Furl N, Begum F, Ferrarese F P, et al. Caricatured facial movements enhance perception of emotional facial expressions. Perception, 2022, 51(5): 313-343

[52] Mery D. True black-box explanation in facial analysis// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 1596-1605

[53] Arriaga O, Valdenegro-Toro M, Plöger P. Real-time convolutional neural networks for emotion and gender classification. arXiv preprint arXiv:1710.07557, 2017

[54] Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421, 2018

[55] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 2921-2929

[56] Xie S, Hu H, Wu Y. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern Recognition, 2019, 92: 177-191

[57] Lucey P, Cohn J F, Kanade T, et al. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. San Francisco. USA, 2010: 94-101

[58] Lyons M, Akamatsu S, Kamachi M, et al. Coding facial expressions with gabor wavelets//Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition. Nara, Japan, 1998: 200-205

[59] Eroglu Erdem C, Turan C, Aydin Z. Baum-2: A multilingual audio-visual affective face database. Multimedia Tools and Applications, 2015, 74(18): 7429-7459

[60] Goodfellow I J, Erhan D, Luc Carrier P, et al. Challenges in representation learning: A report on three machine learning contests. Neural Networks, 2015, 64: 59-63

[61] Dhall A, Goecke R, Lucey S, et al. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark//Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV workshops). Barcelona, Spain, 2011: 2106-2112

[62] Chen Y, Wang J, Chen S, et al. Facial motion prior networks for facial expression recognition//Proceedings of the 2019 IEEE Visual Communications and Image Processing (VCIP), Sydney, Australia, 2019: 1-4

[63] Xue F, Wang Q, Guo G. Transfer: Learning relation-aware facial expression representations with transformers//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 3581-3590

[64] Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA, 2021: 782-791

[65] Zhang Y, Wang C, Ling X, et al. Learn from all: Erasing attention consistency for noisy label facial expression recognition//Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 2022: 418-434

[66] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 618-626

[67] Chattopadhay A, Sarkar A, Howlader P, et al. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks//Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, USA, 2018: 839-847

[68] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450

[69] Wen Z, Lin W, Wang T, et al. Distract your attention: Multi-head cross attention network for facial expression recognition. Biomimetics, 2023, 8(2): 199

[70] Farzaneh A H, Qi X. Facial expression recognition in the wild via deep attentive center loss//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2021: 2402-2411

[71] Jiao Y, Niu Y, Zhang Y, et al. Facial attention based convolutional neural network for 2D+3D facial expression recognition//Proceedings of the 2019 IEEE Visual Communications and Image Processing (VCIP). Sydney, Australia, 2019: 1-4

[72] Jiao Y, Niu Y, Tran T D, et al. 2D+3D facial expression recognition via discriminative dynamic range enhancement and multi-scale learning. arXiv preprint arXiv: 2011.08333, 2020

[73] Chen W, Zhang D, Li M, et al. STCAM: Spatial-temporal and channel attention module for dynamic facial expression recognition. IEEE Transactions on Affective Computing, 2023, 14(1): 800-810

[74] Zhang J, Yu H. Improving the facial expression recognition and its interpretability via generating expression pattern-map. Pattern Recognition, 2022, 129: 108737

[75] Marrero Fernandez P D, Guerrero Pena F A, Ren T, et al. Feratt: Facial expression recognition with attention net//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA, 2019: 1-10

[76] Liu Y, Feng C, Yuan X, et al. Clip-aware expressive feature learning for video-based facial expression recognition. Information Sciences, 2022, 598: 182-195

[77] Ruan D, Yan Y, Chen S, et al. Deep disturbance-disentangled learning for facial expression recognition//Proceedings of the 28th ACM International Conference on Multimedia (MM), Seattle, USA, 2020: 2833-2841

[78] Halawa M, Wöllhaf M, Vellasques E, et al. Learning disentangled expression representations from facial images. arXiv preprint arXiv: 2008.07001, 2020

[79] Jiang J, Deng W. Disentangling identity and pose for facial expression recognition. IEEE Transactions on Affective Computing, 2022, 13(4): 1868-1878

[80] Liu X, Vijaya Kumar B V K, Jia P, et al. Hard negative generation for identity-disentangled facial expression recognition. Pattern Recognition, 2019, 88: 1-12

[81] Liu X, Jin L, Han X, et al. Identity-aware facial expression recognition in compressed video//Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy, 2021: 7508-7514

[82] Teng J, Zhang D, Zou W, et al. Typical facial expression network using a facial feature decoupler and spatial-temporal learning. IEEE Transactions on Affective Computing, 2023, 14(2): 1125-1137

[83] Li H, Yang H, Huang D. Drfer: Learning disentangled

representations for 3D facial expression recognition. arXiv preprint arXiv:2403.08318, 2024

[84] Pantic M, Valstar M, Rademaker R, et al. Web-based database for facial expression analysis//Proceedings of the International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 2005:5

[85] Zhao G, Huang X, Taini M, et al. Facial expression recognition from near-infrared videos. Image and Vision Computing, 2011, 29(9): 607-619

[86] Liu P, Zhou J T, Tsang I W-H, et al. Feature disentangling machine - a novel approach of feature selection and disentangling in facial expression analysis//Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switerland, 2014: 151-166

[87] Xue M, Mian A, Duan X, et al. Learning interpretable expression-sensitive features for 3D dynamic facial expression recognition//Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Lille, France, 2019: 1-7

[88] Li Y, Gao Y, Chen B, et al. Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 3190-3202

[89] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)//Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden, 2018: 2668-2677

[90] Asokan A R, Kumar N, Ragam A V, et al. Interpretability for multimodal emotion recognition using concept activation vectors//Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022: 01-08

[91] Li Z, Wang C, Liu X, et al. Facial expression description and recognition based on fuzzy semantic concepts. Future Generation Computer Systems, 2021, 114: 619-628

[92] Zhang W, Xia M. Pain expression recognition based on SLPP and MKSVM. International Journal of Engineering and Manufacturing, 2011, 1(3): 69-74

[93] Huang W, Zhang W. Driver fatigue recognition based on supervised LPP and MKSVM//Proceedings of the 3rd International Conference on Digital Image Processing (ICDIP 2011). Chengdu, China, 2011: 80091P

[94] Zhou Z-H, Feng J. Deep forest. National Science Review, 2019, 6: 74 - 86

[95] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255

[96] Lin W, Ge Q, Liong S, et al. The design of error-correcting output codes based deep forest for the micro-expression recognition. Applied Intelligence, 2022, 53(3): 3488 - 3504

[97] Kim S, Ko B-C, Nam J. Model simplification of deep random forest for real-time applications of various sensor data. Sensors, 2021, 21(9): 3004

[98] Yeh C-K, Hsieh C-Y, Suggala A S, et al. On the (in)fidelity and sensitivity of explanations//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, USA, 2019: Article 984

[99] Yuan H, Yu H, Gui S, et al. Explainability in graph neural networks: A taxonomic survey. arXiv preprint arXiv: 2012.15445, 2020

[100] Martinez A, Benavente R. The ar face database: Cvc technical report, 1998, 24

[101] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 2019, 28(5): 2439-2450

[102] Thomaz C E, Giraldi G A. A new ranking method for principal components analysis and its application to face image analysis. Image and Vision Computing, 2010, 28(6): 902-913

[103] Valstar M, Pantic M. Induced disgust, happiness and surprise: An addition to the MMI facial expression database//Proceedings of the 3rd International Workshop on Emotion (satellite of LREC): Corpora for Research on Emotion and Affect. Valletta, Malta, 2010: 65-70

[104] Susskind J M, Anderson A K, Hinton G E. The Toronto face database. Department of Computer Science, University of Toronto, Toronto, Canada, Technical Report, 2010, 3: 29

[105] Yin L, Wei X, Sun Y, et al. A 3D facial expression database for facial behavior research//Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 2006: 211-216

[106] Savran A, Alyüz N, Dibeklioğlu H, et al. Bosphorus database for 3D face analysis//Proceedings of the Biometrics and Identity Management: First European Workshop (Bioid 2008). Roskilde, Denmark, 2008: 47-56

[107] Sagonas C, Tzimiropoulos G, Zafeiriou S, et al. 300 faces in-the-wild challenge: The first facial landmark localization challenge//Proceedings of the IEEE International Conference on Computer Vision Workshops. Sydney, Australia, 2013: 397-403

[108] Belhumeur P N, Jacobs D W, Kriegman D J, et al. Localizing parts of faces using a consensus of exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2930-2940

[109] Barsoum E, Zhang C, Ferrer C C, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution//Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo, Japan, 2016: 279-283

[110] Li S, Deng W. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 2584-2593

[111] Mollahosseini A, Hasani B, Mahoor M H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 2017, 10(1): 18-31

[112] Yin L, Chen X, Sun Y, et al. A high-resolution 3D dynamic facial expression database//Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. Amsterdam, The Netherlands, 2008: 1-6

[113] Yan W, Li X, Wang S, et al. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PloS One, 2014, 9(1): e86041

[114] Li X, Pfister T, Huang X, et al. A spontaneous micro-expression database: Inducement, collection and baseline//Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Shanghai, China, 2013: 1-6

[115] Davison A K, Lansley C, Costen N, et al. Samm: A spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing, 2016, 9(1): 116-129

[116] Dhall A, Goecke R, Lucey S, et al. Collecting large, richly annotated facial-expression databases from movies. IEEE MultiMedia, 2012, 19(3): 34-41

[117] Jiang X, Zong Y, Zheng W, et al. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild//Proceedings of the 28th ACM International Conference on Multimedia. New York, USA, 2020: 2881-2889

[118] Wang Y, Sun Y, Huang Y, et al. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 20890-20899

[119] Liu Y, Dai W, Feng C, et al. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal, 2022: 24-32

[120] Ancona M, Ceolini E, Öztireli C, et al. Gradient-based attribution methods, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning: Springer-Verlag, 2022: 169-191

[121] Ji Shou-Ling, Li Jin-Feng, Du Tian-Yu, et al. A survey on techniques, applications and security of machine learning interpretability. Journal of Computer Research and Development, 2019, 56(10): 2071-2096 (in Chinese)
(纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56(10): 2071-2096)

[122] Caruana R, Lou Y, Gehrke J, et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 1721 - 1730

[123] Ruan D, Yan Y, Lai S, et al. Feature decomposition and reconstruction learning for effective facial expression recognition//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA, 2021: 7656-7665

[124] Ruan D, Mo R, Yan Y, et al. Adaptive deep disturbance-disentangled learning for facial expression recognition. International Journal of Computer Vision, 2022, 130(2): 455-477

[125] Wang Y, Zhang X, Hu X, et al. Dynamic network pruning with interpretable layerwise channel selection//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA, 2020: 6299-6306

[126] Yao K, Cao F, Leung Y, et al. Deep neural network compression through interpretability-based filter pruning. Pattern Recognition, 2021, 119: 108056

[127] Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Science Review, 2019, 1(2)

[128] Mollahosseini A, Hasani B, Mahoor M H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 2019, 10(1): 18-31

[129] Du S, Tao Y, Martinez A M. Compound facial expressions of emotion. Proceedings of the National Academy of Sciences, 2014, 111(15): E1454-E1462

[130] Liu Z, Li S, Deng W. Boosting-poof: Boosting part based one vs one feature for facial expression recognition in the wild//Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). Washington, USA, 2017: 967-972

[131] Li S, Deng W. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. International Journal of Computer Vision, 2019, 127(6): 884-906

[132] Plutchik R, Kellerman H. Emotion: theory, research, and experience. London, UK: Psychological Medicine, 1980

[133] Parrott W G. Emotions in social psychology. London, UK: Psychology Press, 2001

[134] Wang W, Sun Q, Chen T, et al. A fine-grained facial expression database for end-to-end multi-pose facial expression recognition. arXiv preprint arXiv:1907.10838, 2019

[135] Shahid A R, Yan H. Squeezexpnet: Dual-stage convolutional neural network for accurate facial expression recognition with attention mechanism. Knowledge-Based Systems, 2023, 269: 110451

[136] Zadeh A B, Liang P P, Poria S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph//Proceedings of the 56th Annual Meeting of the Association-for-Computational-Linguistics (ACL). Melbourne, Australia, 2018: 2236-2246

[137] Tsai Y-H H, Ma M Q, Yang M, et al. Multimodal routing: Improving local and global interpretability of multimodal language analysis//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Virtual, 2020: 1823

[138] Kumar P, Malik S, Raman B. Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. Multimedia Tools and Applications, 2024, 83(10): 28373-28394

[139] Palash M, Bhargava B. Emersk-explainable multimodal emotion recognition with situational knowledge. IEEE

Transactions on Multimedia, 2024, 26: 2785-2794
[140] Kirillov A, Mintun E, Ravi N, et al. Segment anything// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 4015-4026
[141] Bai Y, Geng X, Mangalam K, et al. Sequential modeling enables scalable learning for large vision models. arXiv preprint arXiv:2312.00785, 2023
[142] Fu M, Song Y, Lv J, et al. A versatile framework for analyzing galaxy image data by implanting human-in-the-loop on a large vision model. arXiv preprint arXiv:2405.10890, 2024
[143] Zhang J, Huang J, Jin S, et al. Vision-language models for vision tasks: A survey. arXiv preprint arXiv:2304.00685, 2023
[144] Lu P, Mishra S, Xia T, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 2022, 35: 2507-2521
[145] Yang D, Chen K, Rao J, et al. Tackling vision language tasks through learning inner monologues. arXiv preprint arXiv: 2308.09970, 2023
[146] Yan A, Wang Y, Zhong Y, et al. Learning concise and descriptive attributes for visual recognition//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France,2023: 3090-3100
[147] Yellinek N, Karlinsky L, Giryes R. 3VL: Using trees to teach vision & language models compositional concepts. arXiv preprint arXiv:2312.17345, 2023
[148] Lin B, Xu Y, Bao X, et al. SkinGEN: An explainable dermatology diagnosis-to-generation framework with interactive vision-language models. arXiv preprint arXiv:2404.14755, 2024
[149] Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549, 2023
[150] Schlarmann C, Hein M. On the adversarial robustness of multi-modal foundation models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 3677-3685
[151] Chen F, Han M, Zhao H, et al. X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv:2305.04160, 2023
[152] Zhang H, Li X, Bing L. Video-llama: An instruction-tuned audio-visual language model for video understanding// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Sentosa Gateway. Singapore, 2023: 543-553
[153] Zhao Y, Lin Z, Zhou D, et al. Bubogpt: Enabling visual grounding in multi-modal llms. arXiv preprint arXiv:2307.08581, 2023
[154] Panagopoulou A, Xue L, Yu N, et al. X-InstructBLIP: A framework for aligning X-modal instruction-aware representations to LLMS and emergent cross-modal reasoning. arXiv preprint arXiv:2311.18799, 2023
[155] Tang Z, Yang Z, Khademi M, et al. CoDi-2: In-context, interleaved, and interactive any-to-any generation. arXiv preprint arXiv:2311.18775, 2023
[156] Team G, Anil R, Borgeaud S, et al. Gemini: A family of highly capable multimodal models. arXiv preprint arXiv: 2312.11805, 2023
[157] Lu J, Clark C, Lee S, et al. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. arXiv preprint arXiv:2312.17172, 2023
[158] Lian Z, Sun L, Sun H, et al. Explainable multimodal emotion reasoning. arXiv preprint arXiv:2306.15401, 2023
[159] Yuan J, Sun S, Omeiza D, et al. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. arXiv preprint arXiv:2402.10828, 2024
[160] Richens J G, Lee C M, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nature Communications, 2020, 11(1): 3923
[161] Li Y, Lu G, Li J, et al. Facial expression recognition in the wild using multi-level features and attention mechanisms. IEEE Transactions on Affective Computing, 2020, 14(1): 451-462

ZHANG Miao-Xuan, Ph. D. candidate. Her main research interests include affective computing, facial expression recognition, computer vision, and vision language models.

ZHANG Hong-Gang, Ph. D., associate professor. His research interests include image retrieval, computer vision, and pattern recognition.

Background

Facial expression recognition is an important topic in the field of affective computing and computer vision, with unique practical significance. With the abundance of facial expression datasets and the enhancement of computational power, expression recognition algorithms have gained continuous development.

It is worth noting that while researchers focus on model accuracy, they are also paying increasing attention to the interpretability. Interpretability reflects the extent to which humans can understand the decisions of models. Studies on

interpretability help deepen researchers' understanding of models and ensure their fairness, robustness, and privacy-preserving performance. There have already been several works summarizing the interpretability of machine learning. However, our work systematically summarizes the interpretability in the field of facial expression recognition, classifying it into result interpretability, mechanism interpretability, and model interpretability.

According to the different perspectives and objects of interpretability research, interpretability can be categorized into three main types. Result interpretability refers to the extent to which people with specific expertise can understand the outcomes of the models. Mechanism interpretability focuses on explaining the internal mechanism of models. Model interpretability involves works trying to uncover the decision-making principles of models. Based on the above categorization, this paper reviews of interpretability studies in the field of expression recognition. Specifically, result interpretability in FER mainly includes methods based on text descriptions and the basic structure of the face. Among the mechanism interpretable methods, the attention mechanism in FER is studied, as well as the interpretability methods based on feature decoupling and concept learning. For model interpretability, the focus is primarily on interpretable classification methods. Additionally, the interpretable FER works are compared and analyzed. We acknowledge the existing challenges in current expression recognition interpretability studies. This is followed by a discussion and outlook on future directions, including the interpretability of complex expression recognition, multi-modal emotion recognition, expression and emotion recognition using large models, and the enhancement of generalization ability through interpretability.

Our previous works on FER include "Joint Patch and Multi-label Learning for Facial Action Unit and Holistic Expression Recognition" (TIP), "Deep region and multi-label learning for facial action unit detection" (CVPR), "Multi-label learning with prior knowledge for facial expression analysis" (Neurocomputing), etc.

This work was supported by the National Natural Science Foundation of China No. 62076034. This paper aims to provide interested researchers with a comprehensive review and analysis of the current state of the interpretability of facial expression recognition, and to promote further development in this field.