软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(6):2922−2941 [doi: 10.13328/j.cnki.jos.006497]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

语义关系引导的面部动作单元分析*

李冠彬[1,2], 张锐斐[1], 朱 鑫[1], 林 倞[1]

[1](中山大学 计算机学院, 广东 广州 510006)

[2](人工智能与数字经济广东省实验室 (广州), 广东 广州 510323)

通信作者: 林倞, E-mail: linliang@ieee.org

摘 要: 面部动作单元分析旨在识别人脸图像每个面部动作单元的状态, 可以应用于测谎, 自动驾驶和智能医疗等场景. 近年来, 随着深度学习在计算机视觉领域的普及, 面部动作单元分析逐渐成为人们关注的热点. 面部动作单元分析可以分为面部动作单元检测和面部动作单元强度预测两个不同的任务, 然而现有的主流算法通常只针对其中一个问题. 更重要的是, 这些方法通常只专注于设计更复杂的特征提取模型, 却忽略了面部动作单元之间的语义相关性. 面部动作单元之间往往存在着很强的相互关系, 有效利用这些语义知识进行学习和推理是面部动作单元分析任务的关键. 因此, 通过分析不同人脸面部行为中面部动作单元之间的共生性和互斥性构建了基于面部动作单元关系的知识图谱, 并基于此提出基于语义关系的表征学习算法 (semantic relationship embedded representation learning, SRERL). 在现有公开的面部动作单元检测数据集 (BP4D、DISFA) 和面部动作单元强度预测数据集 (FERA2015、DISFA) 上, SRERL 算法均超越现有最优的算法. 更进一步地, 在 BP4D+数据集上进行泛化性能测试和在 BP4D 数据集上进行遮挡测试, 同样取得当前最优的性能.

关键词: 面部动作单元分析; 深度学习; 计算机视觉

中图法分类号: TP391

中文引用格式: 李冠彬, 张锐斐, 朱鑫, 林倞. 语义关系引导的面部动作单元分析. 软件学报, 2023, 34(6): 2922–2941. http://www.jos.org.cn/1000-9825/6497.htm

英文引用格式: Li GB, Zhang RF, Zhu X, Lin L. Semantic Relationships Guided Facial Action Unit Analysis. Ruan Jian Xue Bao/Journal of Software, 2023, 34(6): 2922–2941 (in Chinese). http://www.jos.org.cn/1000-9825/6497.htm

Semantic Relationships Guided Facial Action Unit Analysis

LI Guan-Bin[1,2], ZHANG Rui-Fei[1], ZHU Xin[1], LIN Liang[1]

[1](School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China)

[2](Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), Guangzhou 510323, China)

Abstract: The main purpose of facial action unit analysis is to identify the state of each facial action unit, which can be applied to many scenarios such as lie detection, autonomous driving, intelligent medical, and others. In recent years, with the popularization of deep learning in the field of computer vision, facial action unit analysis has attracted extensive attention. Face action unit analysis can be divided into two different tasks: face action unit recognition and face action unit intensity estimation. However, the existing studies usually only address one of the problems. More importantly, these methods usually only focus on designing or learning complex feature representations, but ignore the semantic correlation between facial action units. Actually, facial action units often have strong interrelationships. How to effectively use semantic knowledge for learning and reasoning is the key to facial action unit analysis tasks. This study explores to model the semantic relationship of facial action units by analyzing the symbiosis and mutual exclusion of AUs in various facial behaviors and organize the facial AUs in the form of structured knowledge-graph, and then propose an AU semantic relationship embedded representation learning (SRERL) framework. The experiments are conducted on three benchmarks: BP4D, DISFA,

* 基金项目: 国家自然科学基金 (61976250, U1811463); 广东省基础与应用基础研究基金 (2020B1515020048)
收稿时间: 2021-03-26; 修改时间: 2021-06-07, 2021-07-08, 2021-08-17; 采用时间: 2021-09-17; jos 在线出版时间: 2022-11-30
CNKI 网络首发时间: 2022-12-01

and FERA2015 for both facial action unit analysis tasks. The experimental results show that the proposed method outperforms the previous work and achieves state-of-the-art performance. Furthermore, the experiments are also conducted on the BP4D+ dataset and occlusion evaluation is performed on the BP4D dataset to demonstrate the outstanding generalization and robustness of proposed method.

Key words: facial action unit analysis; deep learning; computer vision

人脸面部表情及其行为是个人传递情感的重要渠道之一. 智能化的面部表情分析在人机交互, 智能教育, 智慧医疗等计算机视觉任务中存在巨大应用价值, 近年来吸引了越来越多的研究兴趣. 目前, 用于测量和描述面部行为的最通用的方法是由 Ekman 等人[1]在 1978 年提出的面部动作编码系统 (facial action coding system, FACS) 中所定义的面部动作单位 (AUs), 描述人脸局部区域的动作变化, 受人脸面部肌肉所控制. 相比于表情类别, AU 是描述人脸面部行为更细微, 更客观的一种方式.

面部动作单元分析问题具体可以分为 AU 检测和 AU 强度预测两个不同的任务. AU 检测的主要目的是检测输入人脸图像的每个 AU 的状态, 包含激活和未激活两种, 属于单帧图像的多标签二分类问题; AU 强度预测是指预测输入图像每个 AU 的强度, 是多标签回归问题. 相比于 AU 检测只判断状态是否激活, AU 强度预测具有更丰富和精细的类别标签, 可以进一步反映出激活的强度等级. 早期的 AU 分析传统方法大多专注于设计更具区分性的手工特征 (例如形状或外观特征) 或更有效的区分性学习方法[2−4]. 近年来, 深度卷积神经网络因其强大的特征表示能力和端到端高效学习方案而被广泛用于 AU 分析中, 极大地促进了该领域的发展[5−8]. 然而, 最近基于深度卷积神经网络的研究无一例外地致力于设计更深, 更复杂的网络结构, 以数据驱动的方式学习更强大的特征表示, 而没有明确考虑和建模面部器官的局部特性和相互关系. 通常情况, 面部动作单元并不是彼此独立的. 如图 1 所示, 高兴表情会同时激活 AU6 和 AU12, 因为当我们微笑时, 我们通常会“举起脸颊”和“拉起唇角”. 另一方面, 由于面部结构限制, 某些动作单元通常不太可能同时激活, 例如我们不能同时张开嘴巴和抬起脸颊.

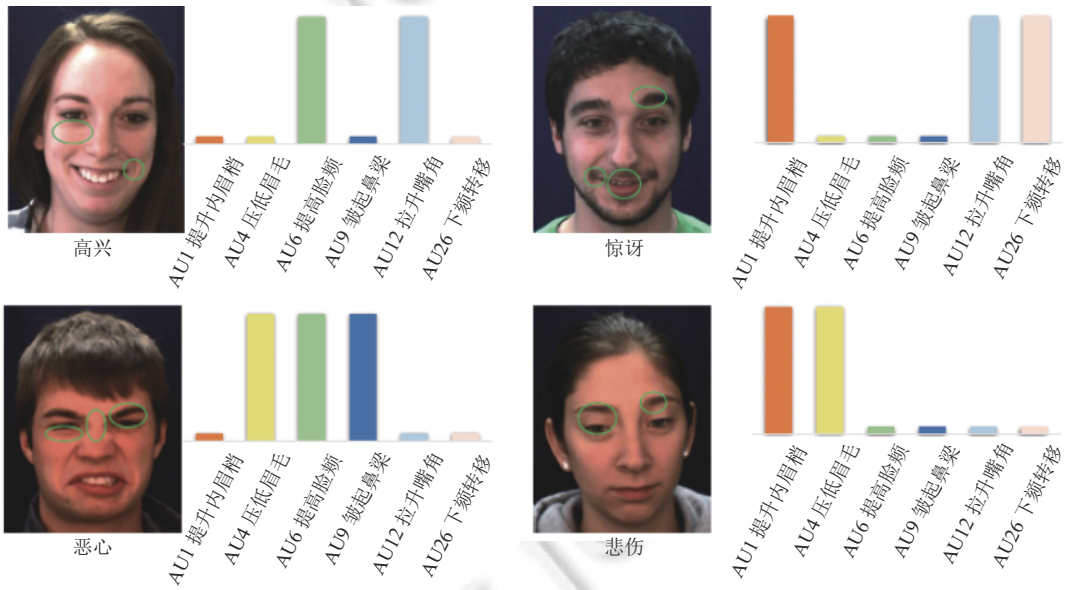


图 1　不同表情下面部动作单元状态示意图

考虑到上述关系, 一些研究工作采用对 AU 关系进行建模来提高 AU 分析精度. 例如动态贝叶斯网络 (DBN)[9]和受限玻尔兹曼机 (RBM)[10]等. 然而, 现有基于 AU 关系的分析方法具有以下 3 个缺点.

1) 现有基于 AU 关系的模型大多是基于低级手工特征设计的, 并通常作为一种后期处理方式嵌入到复杂的分类模型中, 独立于特征学习阶段, 因此限制了特征提取的性能.

2) 现有方法通常只针对 AU 检测或者 AU 强度预测其中一个问题, 该方法基于观察有限的面部表情来捕获成对 AU 之间的局部依赖性, 并且这些成对数据没有组合起来形成用于更全面的 AU 关系推理的图形结构, 因此不

具有推广性和迁移性.

3) 由于现有的 AU 关系建模依赖于预先的特征提取, 因此整个算法框架无法端到端地运行, 这极大地限制了模型的效率和性能.

鉴于上述缺点, 并受到图神经网络的可微性及其在关系学习中优越性能的启发, 本文提出了一种 AU 语义关系引导表征学习的框架 (semantic relationship embedded representation learning, SRERL), 通过充分利用 AU 之间的关系来引导模型学习更具有区分性的特征. 具体地说, 本文采用结构化的知识图谱对 AU 关系进行建模, 相比于之前的模型, 形成了更为全面的 AU 关系推理的图形结构, 并将门控图神经网络 (GGNN)[11]集成到基于注意力机制的多尺度特征提取模块中, 以端到端的方式运行, 提高了模型的整体学习能力和效率, 有力缓解了现有 AU 关系建模方法的缺点和不足. 由于算法学习到的特征同时融合了表观信息和 AU 关系推理, 本文提出的模型能应对更复杂的场景, 比如光照变换和人脸遮挡等, 也更具有鲁棒性.

总而言之, 本文的主要贡献如下.

1) 本文研究了如何对面部动作单元语义关系进行建模, 通过分析不同人脸面部行为中面部动作单元被激活的规律总结面部动作单元之间的关系, 并在 AU 检测和 AU 强度预测两种不同任务条件下构建基于 AU 语义关系的知识图谱. 本文提出的 AU 知识图谱具有良好的迁移性和推广性, 可以应用于不同面部动作单元数据集.

2) 本文提出的算法有效结合了卷积神经网络和图神经网络的优点, 通过 AU 之间的语义关系传播增强对应人脸区域的特征表示, 并且能同时被应用到 AU 检测和 AU 强度预测两个不同的任务中. 这种方法利用 AU 语义关系引导模型学习到更具有区分性的特征, 使得特征同时融合了表观信息和 AU 关系推理, 在更为复杂的场景下, 例如光照变换和人脸遮挡等, 可有效地利用可见区域的 AU 情况和 AU 之间的语义关系来引导不可见区域 AU 的预测, 进而提升了算法的鲁棒性.

3) 本文在现有公开的两个面部动作单元检测数据集 (BP4D、DISFA) 和两个面部动作单元强度预测数据集 (FERA2015、DISFA) 上进行实验验证. 实验结果表明, SRERL 算法在上述两个面部动作单元分析任务中均超越现有最优的算法. 更进一步, 本文在 BP4D+数据集上进行泛化性能测试和在 BP4D 数据集上进行遮挡测试, 同样取得当前最优的性能. 同时本文通过可视化分析模型的可解释性和通过消融实验验证各个模块的合理有效性.

1 面部动作单元关系建模

面部动作单元由人脸面部肌肉所控制, 受限于人脸的结构性, 面部动作单元之间往往存在着一定的关系. Corneanu 等人[12]通过结构化推理模块分析得出 AU4 和 AU17 具有负相关性, Du 等人[13]认为在 90% 以上的时候, 微笑这种面部行为会同时激活拉升嘴角 AU12 和提高脸颊 AU6 这两个面部动作单元. 通过总结不同人脸面部行为中 AU 被激活的规律, 本节提出了面部动作单元检测和面部动作单元强度预测两种不同任务条件下 AU 建模的方法.

1.1 面部动作单元检测

在面部动作单元检测任务中, 每个 AU 只有激活和未被激活两种状态, 属于多标签二分类问题. 图 1 列举了不同表情下面部动作单元的状态, 可以看到面部动作单元之间并不是彼此独立的, 不同的面部动作单元通过组合可以构成不同的人脸表情, 比如恶心可以由 AU4 (压低眉毛)、AU6 (提高脸颊) 和 AU9 (皱起鼻梁) 组成, 悲伤可以由 AU1 (提升内眉梢) 和 AU4 (压低眉毛) 组成等. 本文借鉴以往工作对 AU 关系的研究[12,14], 将 AU 之间的关系定义为共生性和互斥性两种.

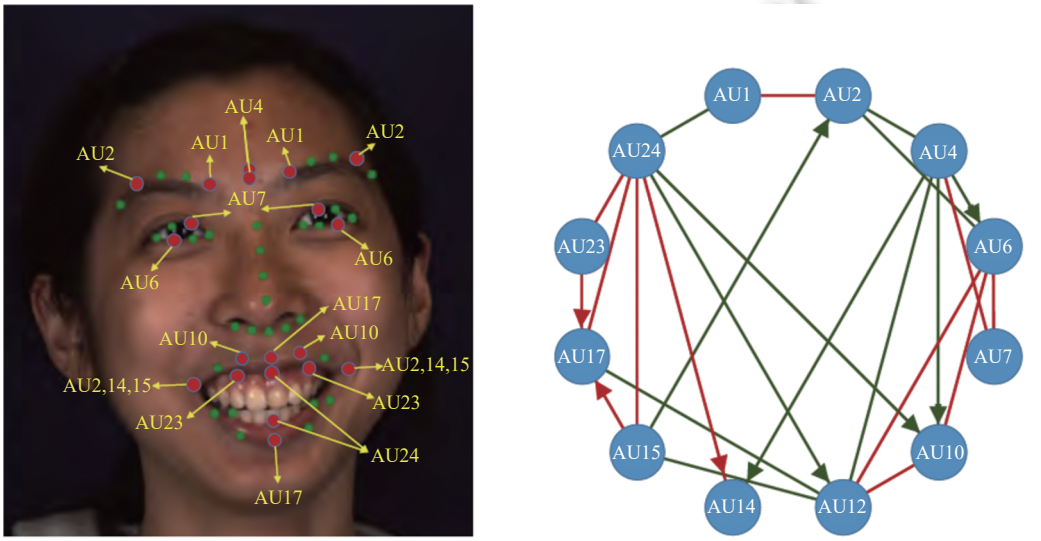
共生关系是指某个 AU 的激活往往伴随着另一个 AU 的激活, 例如由于脸部肌肉和嘴部肌肉的相互影响, AU6 提高脸颊和 AU12 拉升嘴角往往会同时被激活, 因此这两者之间的条件概率 P(AU12|AU6) 和 P(AU6|AU12) 的值都会很高.

互斥关系是指某两个 AU 很少会同时出现, 比如在自然的面部情绪下, AU4 压低眉毛和 AU12 拉伸嘴角几乎不会被同时激活, 因此 P(AU4|AU12) 和 P(AU12|AU4) 的值都很低. 考虑到条件概率的局限性和 AU 被激活的频繁程度不同, 本文通过统计两两 AU 之间的条件概率与其自身被激活的概率对 AU 关系进行建模, 具体公式如公式 (1)

所示:

$$a_{i,j} = P(y_i = 1 | y_j = 1) - P(y_i = 1) \tag{1}$$

其中, y_n指的是第 n 个 AU 的标签, $a_{i,j} \in A$, A 表示 AU 关系抽象成的知识图谱, 并以矩阵的方式表现, 由条件概率与自身概率的差构成, 数值大小代表了 AU 之间关系的强弱, 而正负则分别表示共生和互斥关系. 由于条件概率的不对称性, 与现有其他 AU 建模方法[12,14]不同, 本文构建的 AU 知识图谱属于有向图, 如图 2(b) 所示, 其中绿线代表共生关系, 红线代表互斥关系, 带箭头的线代表单向关系, 无箭头的线代表双向关系. 图 2(a) 为基于人脸关键点定义 AU 中心位置. 绿色点代表人脸关键点, 红色点代表 AU 中心位置. 其中, 存在位置一致但代表的动作含义不同的情况, 比如 AU12, AU14 和 AU15.



(a) 基于人脸关键点定义 AU 中心位置　　(b) 构建的 AU 关系图

图 2　基于人脸关键点定义 AU 中心位置和构建的 AU 关系图

1.2　面部动作单元强度预测

AU 强度预测是 AU 检测的扩展任务, 相比于 AU 检测中简单地将 AU 分为激活和未激活两种状态, AU 强度预测具有更丰富的标签信息. 在 AU 强度预测任务中, 将 AU 强度分为 0–5 这 6 种等级, 其中 0 表示未被激活状态, 1–5 则表示随着数值的增加 AU 激活强度逐渐加大. 由于人脸肌肉的结构性, AU 之间的关系不仅可以体现激活和未激活这两种状态的共生互斥性上, 不同 AU 的强度之间也会相互影响. 比如 AU1 内侧眉毛上扬和 AU2 外侧眉毛上扬都是由人脸额肌所控制, 人们很难只激活 AU1 而不激活 AU2, 同样当 AU1 强度的提升势必会带来 AU2 强度的提升; 而 AU2 外侧眉毛上扬和 AU4 眉毛下压分别由不同肌肉所控制, 从其含义就可以看出这两个 AU 很难被同时激活, 当其中某个 AU 强度增大时, 必然会导致另一个 AU 处于未激活状态. 图 3 展示了这一关系, 可以看到当 AU12 (拉升嘴角) 的强度到达 2 时, AU10 (提升上嘴唇) 和 AU14 (压出梨涡) 同时被激活, 其强度随着 AU12 的强度增大而增大.

本文尝试采用皮尔逊相关系数度量两两 AU 强度之间的相互关系, 具体公式如公式 (2) 所示, 其中 X, Y 代表不同的 AU 编号, σ_X 为样本 X 的标准差, μ_X 为样本 X 的均值, 通过计算两个 AU 之间的协方差与标准差的商便可以得到最终皮尔逊相关系数的值.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{2}$$

公式 (2) 得到的结果位于 [−1, 1] 之间, 其中越接近 1 代表共生关系越强, 反之越接近−1 则代表互斥关系越强, 0 表示这两个 AU 之间不存在线性关系.

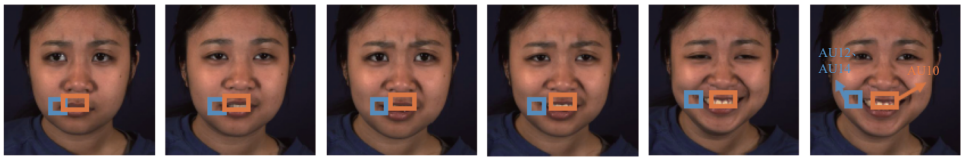


图 3 AU 强度变化示意图

2 算法框架设计与实现

针对面部动作单元分析问题, 本文提出了基于语义关系的表征学习框架 (SRERL), 本节将从框架设计到具体实现细节对模型进行详细介绍, 包括模型网络结构设计, 损失函数设计以及模型的训练方式等.

2.1 算法整体框架

SRERL 是一个基于图神经网络和卷积神经网络的面部动作单元分析模型, 通过 AU 之间的语义关系传播增强对应人脸区域的特征表示, 并且能同时被应用到 AU 检测和 AU 强度预测两个不同的任务中.

如图 4 所示, SRERL 框架主要由 3 部分组成: (1) 多尺度的主干神经网络模块, 用以提取多尺度的全局特征; (2) 基于注意力机制的区域学习模块, 通过利用面部动作单元的位置, 在全局特征图上提取局部特征并输入参数不共享的区域学习通道以获得自适应的局部特征; (3) 基于语义关系的图神经网络模块, 利用面部动作单元之间的语义关系引导模型学习更具有区分性的特征. 算法 1 展示整个框架的流程, 输入一批带有人脸的图像 I 和利用第 1 节中 AU 关系建模算法构建的基于 AU 语义关系的知识图谱 A, 首先经过数据预处理进行人脸对齐并计算面部关键点 M, 利用关键点信息获取如图 2 所示的 AU 中心位置 P. 接着将对齐后的人脸图像输入多尺度主干神经网络模块 $\mathcal{N}_{\text{global}}$ 得到多尺度的人脸全局特征图 f_g, 然后利用 AU 中心位置 P 可以从特征图 f_g 上提取每个 AU 的局部特征图输入区域学习模块 $\mathcal{N}_{\text{local}}$ 得到自适应的 AU 局部特征 f_l, 最后将每个 AU 的局部特征 f_l 输入图神经网络模块 $\mathcal{N}_{\text{GGNN}}$ 对应节点, 通过图网络传播算法实现基于语义关系 A 的特征传播, 实现 AU 状态分析并输出结果 O. 整个框架有效结合了卷积神经网络卓越的提取特征能力和图神经网络对关系推理的能力, 可以端到端高效率运行, 并适用 AU 检测和 AU 强度预测两种不同的任务.

算法 1. SRERL 框架流程.

Input: 输入图像 $I=i_1,\ldots,i_N$; AU 关系知识图谱 A;

Output: 面部动作单元的状态 $O=o_1,\ldots,o_N$.

1. for k in $[1,N]$ do
2. $\quad i_k^{\text{aligned}}, m_k=\mathcal{T}(i_k)$ ◇人脸对齐并计算面部关键点
3. end for
4. $P=\mathcal{L}(M)$ ◇ 计算 AU 中心位置
5. for k in $[1,N]$ do
6. $\quad f\_g_k=\mathcal{N}_{\text{global}}(i_k^{\text{aligned}})$ ◇ 提取全局特征
7. $\quad$ for j in $[1,C]$ do
8. $\quad\quad f\_c_k^j=crop(f\_g_k,p_k^j)$ ◇ 提取局部特征
9. $\quad\quad f\_l_k^j=\mathcal{N}_{\text{local}}^j(f\_c_k^j)$
10. $\quad$ end for
11. $\quad o_k=\mathcal{N}_{\text{GGNN}}(f\_l_k^1,\ldots,f\_l_k^C;A)$
12. end for

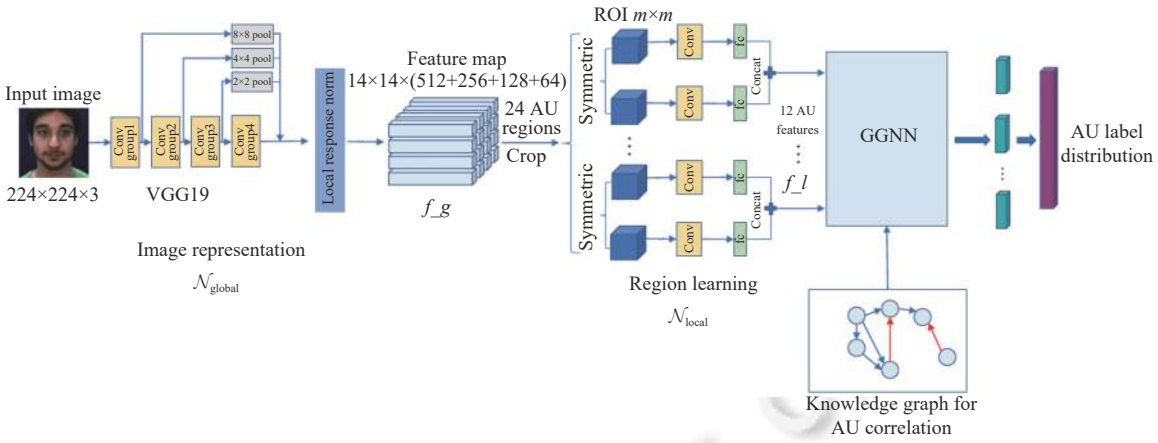


图 4 SRERL 整体框架示意图

2.2 网络结构设计

2.2.1 多尺度主干神经网络模块

受到 VGG 模型在人脸识别[15]中取得优越性能的启发, 本文采用 VGG19 网络[16]作为框架的主干神经网络. VGG19 按池化层可以分为 5 个组, 出于对框架性能和分辨率的综合考虑, 本文使用前 4 组卷积作为框架的初始全局特征提取器. 输入图像大小为 224×224, 经前 4 组卷积的前向传播输出 14×14 的全局特征图. 然而面部动作单元对应于人脸所占的区域大小并不是统一的, 比如 AU6 (提高脸颊) 所对应的区域远比 AU12 (拉升嘴角) 更大, 单一尺度的特征图无法满足人脸所有面部动作单元特征的提取. 同时受到 DenseNet[17]特征复用和目标检测中特征金字塔[18]的应用, 本文设计了一种基于特征复用的多尺度特征提取方式.

$$multi\_scale\_feature = Concat(P_8(F_1), P_4(F_2), P_2(F_3), F_4) \tag{3}$$

其中, $Concat$ 表示连接, P_i 代表步长和核大小为 i 的最大池化层, F_j 代表第 j 组的特征图. 通过将每个组的特征图使用最大池化层缩小到与最后一组特征图相同大小, 再将这 4 组特征图按通道这个维度连接起来形成 $14\times14\times(512+256+128+64)$ 大小的多尺度全局特征图, 由于池化和连接操作并不存在任何参数, 多尺度特征图的形成并不会引入额外的模型参数而导致模型复杂度的提升. 接着将该特征图输入局部响应层进行归一化, 具体如公式 (4) 所示:

$$b_c = \frac{a_c}{\left(k + \dfrac{\alpha\times C}{n}\sum_{c'=\max(0,c-n/2)}^{\min(C-1,c+n/2)} a_{c'}^2\right)^{\beta}} \tag{4}$$

其中, k, α, β 为超参数, C 表示特征图的通道数, a_c 则代表特征图上位于第 c 通道的一个像素, n 表示需要归一化的相邻通道数目, 该层通过在特征图通道之间进行局部归一化有助于是整个框架更好的收敛, 并将归一化后的特征图输入框架之后的模块.

2.2.2 基于注意力机制的区域学习模块

由于人脸结构具有复杂性, 其特征不仅包括面部行为, 也包括长相, 年龄等一些与 AU 分析任务无关的信息, 因此去除冗余特征对 AU 分析来说尤为重要. 本文采用基于注意力机制的局部特征提取方法并结合区域学习通道, 得到更具有自适应的 AU 局部特征.

与 Lin 等人[18]相似, 本文在主干神经网络得到的多尺度特征图上以每个 AU 位置为中心截取 $m\times m$ 大小的局部特征图作为自适应区域学习通道的输入. 区域学习通道是为每个 AU 单独提取局部特征所设计的通道, 假设有 N 个 AU, 由于人脸的对称性, 共抽取 $2N$ 块 $m\times m\times(512+256+128+64)$ 大小的局部特征, 因此整个区域学习通道就有 $2N$ 个分支, 每个分支接有一个卷积核 3×3 大小的卷积层和一个输出维度为 150 的全连接层, 卷积层和全连

接层均采用 ReLU 作为激活函数. 区别于传统 CNN 共享卷积核的方式, 区域学习通道中每条分支的卷积层和全连接层参数都是独立的, 目的是为了能学习到更适应于该 AU 的局部特征.

2.2.3 基于语义关系的图神经网络模块

受到图神经网络[19,20]在关系推理任务中优越性能的启发, 本文尝试采用 GGNN[11]作为 SRERL 框架的语义关系推理模块. 通过第 1 节提及的 AU 关系建模方法构建成基于 AU 语义关系的知识图谱, 该知识图谱由节点集 V 和邻接矩阵 A 组成. 如图 4 所示, 通过主干网络和区域学习这两个模块可以得到 $2N$ 个 AU 局部特征, 定义 f_i 为第 i 块区域的特征, 其中 $i \in [0,1,\ldots,2N-1]$. 考虑到人脸的对称性以及知识图谱中每个 AU 对应一个节点的先验知识, 我们可以得到每个节点的输入特征为 $x_v = concat(f_{2v}, f_v)$, 通过将输入特征 x_v 补零实现对每一个节点 $v \in V$ 的初始化, 即 $h_v^{(0)} = [x_v^T, 0]$, 表示 v 节点 $t=0$ 时刻的初始状态. 接着实现整个图网络的传播, 公式 (5) 表示 AU 语义关系 A 引导下的特征交互传播, 公式 (6) 类似于 GRU[21]的构造, 通过门控的方式实现特征的更新.

$$a_v^{(t)} = A_v^T[h_1^{(t-1)} \ldots h_{|V|}^{(t-1)}]^T + b \tag{5}$$

$$\begin{cases} z_v^{(t)} = \sigma(W^z a_v^{(t)} + U^z h_v^{(t-1)}) \\ r_v^t = \sigma(W^r a_v^{(t)} + U^r h_v^{(t-1)}) \\ \widetilde{h}_v^t = \tanh(W a_v^{(t)} + U(r_v^t \odot h_v^{(t-1)})) \\ h_v^{(t)} = (1 - z_v^t) \odot h_v^{(t-1)} + z_v^t \odot \widetilde{h}_v^t \end{cases} \tag{6}$$

其中, tanh 和 σ 分别为双曲正切激活函数和逻辑激活函数, W 、 U 为需要学习的超参数, 通过结合节点当前状态和该节点的邻接信息实现对该节点的一轮特征传播更新. 在经过 T 轮传播以后, 可以得到每个节点最终的状态. 由于本框架的最终目的是分析每个面部动作单元的状态, 即节点级别的预测, 本文尝试将最终状态与初始状态相连接, 并接以全连接层进行预测, 具体如公式 (7):

$$o^v = g(h_v^{(T)}, x_v) \tag{7}$$

其中, g 为全连接层, x_v 为每个节点的原始输入特征, $h_v^{(T)}$ 为每个节点 T 轮传播后的特征, 最后将全连接层得到的结果输入 Sigmoid 激活函数便得到每个 AU 的分析结果.

2.3 损失函数设计

本文提出的 SRERL 框架可以同时被应用到 AU 检测和 AU 强度预测两种不同的任务中, 本节将具体介绍这两种不同任务条件下损失函数的具体设计.

2.3.1 面部动作单元检测

检测的任务是识别输入图像上每个面部动作单元是否被激活, 属于多标签二分类问题. 数据不平衡是 AU 检测中的一个常见问题, 尤其是在多标签训练时候, 无法通过简单有效的过采样或者欠采样方法来实现标签平衡, 而不平衡的标签训练会严重降低模型的精度.

Li 等人[6]设计了一套复杂的采样算法试图平衡样本标签的正负样本比例, 然而由于数据的多样性有限, 该采样算法并不能带来额外的多样性提升, 因此算法的性能仍旧不太理想. 本文尝试从另一种角度出发, 提出了一种加权损失函数, 即通过给正负样本乘以对应的权重来解决标签不平衡问题. 由于 AU 检测是典型的分类问题, 本文采用交叉信息熵作为基础损失函数, 同时加上平滑项用于减少异常样本带来的影响, 最后在交叉信息熵的两端分别乘以正负样本比例进行加权, 具体如公式 (8) 所示:

$$loss = -\frac{1}{C \cdot N} \sum_{j=1}^{N} \sum_{i=1}^{C} 2 \left\{ r_{\text{neg}}^i [l_j^i = 1] \cdot \log\left(\frac{p_j^i + 0.05}{1.05} \right) + r_{\text{pos}}^i [l_j^i = 0] \log\left(\frac{1.05 - p_j^i}{1.05} \right) \right\} \tag{8}$$

其中, l 代表样本标签, p 代表模型预测的结果, $[x]$ 为判断函数, 当且仅当等式 x 为真时为 1, 否则为 0. N 为模型处理图片的批大小, C 代表 AU 的数目, r_{pos}^i 和 r_{neg}^i 为常数, 可以由公式 (9) 计算得到, 即通过统计训练集中每个 AU 的正负样本数目再除以总的样本数目 M .

$$\begin{cases} r^i_{\text{pos}} = \dfrac{\sum_{k=1}^{M}[l^i_k = 1]}{M} \\ r^i_{\text{neg}} = 1 - r^i_{\text{pos}} \end{cases} \tag{9}$$

2.3.2 面部动作单元强度预测

AU 强度预测主要目的是预测输入图像上每个面部动作单元被激活的强度. 相比于 AU 检测任务, AU 强度预测任务更具有挑战性. AU 强度预测范围属于 [0,5], 而本文提出的框架 SRERL 在经过最后的 Sigmoid 层后输出结果范围位于 [0,1], 为了匹配 AU 强度预测任务, 本文通过对 SRERL 框架第 i 个 AU 的预测乘以 5 得到最终 AU 强度的预测结果.

针对 AU 强度回归问题, 本文采用均方误差损失函数 (MSE) 作为基础目标函数. 与 AU 检测任务类似, 在 AU 强度回归任务中存在着更严重的数据不平衡问题, 为此本文提出了基于均方误差的加权损失函数, 形式如公式 (10) 所示:

$$loss = -\frac{1}{C \cdot N}\sum_{j=1}^{N}\sum_{i=1}^{C}(1 - r^i_{l^i_j})(l^i_j - p^i_j)^2 \tag{10}$$

其中, N 为批样本数量, C 为需要预测的 AU 数目, l^i_j 为第 j 个样本第 i 个 AU 的标签, 而 p^i_j 则为对应的预测值. $r^i_{l^i_j}$ 由公式 (11) 计算得到, 代表该样本对应 AU 的强度占总样本的概率.

$$r^i_l = \frac{\sum_{k=1}^{M}[l^i_k == l]}{M} \tag{11}$$

2.4 训练方法

考虑到模型参数复杂而现有 AU 数据集多样性不足, 为了防止模型过拟合, 本文采用多阶段学习策略进行训练, 主要分为以下 3 个步骤.

阶段 1. 微调 ImageNet[22]上训练好的 VGG19 模型, 将 VGG19 模型的前 4 组参数作为 SRERL 框架中多尺度主干神经网络模块的参数.

阶段 2. 对多尺度主干神经网络模块得到的特征图提取局部特征作为区域学习模块的输入, 训练每个区域对应通道的参数.

阶段 3. 固定主干神经网络模块和区域学习模块的参数, 训练图神经网络模块. 以上 3 个阶段的训练均采用加权损失函数作为监督信息, 指导整个模块的参数学习.

在每个阶段训练时, 取在验证集上具有最优性能的模型作为下一阶段的输入特征提取器.

3 实验结果和分析

为了验证 SRERL 算法的有效性, 本文在现有的面部动作单元检测数据集 (DISFA 和 BP4D) 和面部动作单元强度预测数据集 (FERA2015 和 DISFA) 上进行实验, 比较与其他先进算法的优劣. 为了验证算法的泛化性能和推广性, 本文将采用跨数据集测试方式, 并取得了卓越的效果. 同时为了验证本文提出的算法能应对面部遮挡场景, 本文在常用面部动作单元检测数据集上进行遮挡性能测试, 并与现有方法进行公平比较. 除此之外, 本文还通过消融实验验证了本文算法各个模块的有效合理性, 并进一步探索了网络的可解释性.

3.1 实验设置

3.1.1 数据集和评价指标

本文在以下 4 个公开数据集上对算法进行验证: BP4D[23], DIFA[24], FERA2015[25]和 BP4D+[26], 其中 BP4D、DISFA 和 BP4D+数据集用于面部动作单元检测, FERA2015、DISFA、BP4D+数据集用于面部动作单元强度预测. 这 4 个数据集皆由领域专家基于面部动作编码系统 (FACS)[1]进行标注.

本文在面部动作单元检测和面部动作单元强度预测两个不同任务下进行实验: 针对面部动作单元检测任务, 本文采用 $F1$ 和 AUC 作为评价指标; 针对面部动作单元强度预测任务, 本文采用组内相关系数 (ICC(3, 1))[27]、均方误差 (MSE) 和绝对误差 (MAE) 作为评价指标.

3.1.2 实验参数设置

为了验证本文提出 SRERL 模型的有效性, 本文在操作系统为 Ubuntu 16.04 和 GPU 为 12 GB 显存的 NVIDIA GeForce GTX TITAN X 服务器上进行实验. 本文采用 OpenCV[28]进行数据预处理和 dlib 库[29]实现人脸关键点检测. 本文提及的所有网络结构实现都是基于开源深度学习框架 PyTorch[30]. 针对模型超参数, 通过权衡考虑模型精度和模型大小, 本文将数据预处理部分输入图像缩放大小 N 设为 224. 局部响应归一化层 LRN3-5 中 α 设为 0.002、α 设为 0.75、k 设为 2. 区域学习模块提取局部特征图部分, 局部特征图大小 M 设为 6. 图神经网络模块图的节点数目依据数据集标注 AU 个数所决定.

在网络优化策略上, 本文采取 Adam[31]作为优化器, 以学习率 0.000 1 和批图像大小 64 为设定进行网络参数学习. 在测试阶段, 本文采用阈值为 0.5 对模型预测结果二值化进行 $F1$ 分数计算. 同时在数据预处理阶段, 由于存在极少数的图像因为较大的头部姿势偏转或位移导致图像只存在部分面部信息, 而无法进行人脸检测或面部关键点识别使得人脸对齐, 因此本文在实验部分并没有将这样的样本考虑在内. 而 FERA2015 数据集也存在部分样本某些 AU 强度标注丢失, 本文在同样在实验中也忽略这些样本. 由于数据集规模较大 (十几万张图片), 减去少量的特殊样本 (几百张图片) 并不会影响实验的公平性.

3.2 面部动作单元分析结果比较

3.2.1 面部动作单元检测

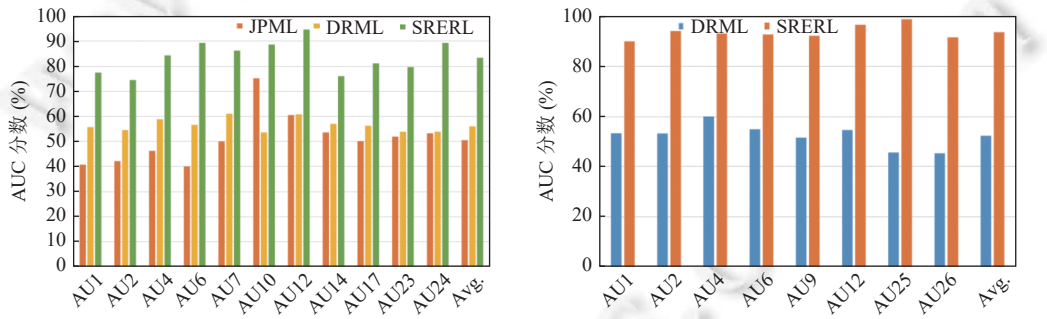
针对面部动作单元检测任务, 本文与近 5 年来先进的 AU 检测算法进行比较: JPML[3], DRML[5], EAC[32], DSIN[12], JAA[33], ARL[34], LP-Net[35]. 表 1 和表 2 分别展示了 BP4D 数据集和 DISFA 数据集中 $F1$ 分数比较, 其中加粗代表最优结果, [] 代表次优结果, 其对应 AUC 结果展示在图 5 中. 分析可以发现, 本文提出的面部动作单元分析框架在 AU 检测任务中优于当前所有先进的方法. 与 JPML 方法相比, SRERL 在 BP4D 数据集上的 $F1$ 分数和 AUC 分别提高了 39.4% 和 65.1%, 主要原因是 JPML 算法基于手工设计的特征而不是端到端的可训练模型. DRML, EAC, DSIN, JAA, ARL 和 LP-Net 都是基于当前流行的深度卷积神经网络的模型, 相比于 DRML 方法, 本文提出的 SRERL 模型在 BP4D 数据集上的 $F1$ 分数和 AUC 分别提高了 32.5% 和 48.9%, 在 DISFA 数据集上的 $F1$ 分数和 AUC 分别提高了 119.5% 和 79.0%, 其性能提升的主要原因在于: 相比于本文提出的模型, DRML 有着更浅的神经网络结构, 而 SRERL 利用了注意力机制实现自适应局部特征, 同时 DRML 并没有对 AU 之间关系进行建模. 就 $F1$ 分数而言, 本文提出的 SRERL 模型在 BP4D 数据集上相比于 EAC 模型和 JAA 模型分别提高了 14.5% 和 6.7%, 在 DISFA 数据集上分别提高了 20.8% 和 4.6%, 其主要原因在于这两种方法都采用了注意力机制而忽略了 AU 之间关系, 仅采用全连接层去隐式的学习这种关系, EAC 方法通过定义 AU 中心位置提取局部特征并预测结果, JAA 方法同时执行面部关键点预测和面部动作单元检测, 并利用模型预测的关键点辅助面部动作单元的特征学习. DSIN, ARL, LP-Net 都是基于 AU 关系建模的方法, 其中 DSIN 在统一框架中同时对深度特征学习和结构化 AU 关系进行了建模. 但是, DSIN 的关系推断部分作用在标签级别, 是一种后处理方式, 并且与特征表示隔离. 本文提出的 SRERL 在 BP4D 和 DISFA 上的 $F1$ 分数分别比 DSIN 分别高 8.7% 和 9.3%, 这很好地证明了特征表达和语义关系建模的联合优化对于面部动作单元检测的重要性, 同时也证明了图神经网络在 AU 关系推理中的有效性. LP-Net 采用 LSTM 对 AU 关键进行建模, 并利用人脸关键点去除面部长相差异带来的影响, 但是由于其忽略了面部关键点对于区域特征提取的重要性而只是粗糙地将特征图的每一格作为局部区域输入 LSTM. 与之相比, SRERL 在 BP4D 和 DISFA 数据集上的 $F1$ 分数分别提高了 4.9% 和 3.0%. ARL 同时利用了空间和通道上的注意力机制, 并利用了条件随机场学习特征图上每个像素之间的关系, SRERL 相比 ARL 方法在 DISFA 数据集上 $F1$ 分数下降了 0.2%, 但在 BP4D 数据集上提高了 4.7%. 综上分析, 通过与当前先进的算法比较, 充分论证了本文提出的 SRERL 方法在面部动作单元检测任务中的有效性.

表 1 BP4D 数据集上 *F*1 分数比较 (%)

AU	JPML	DRML	EAC	DSIN	JAA	ARL	LP-Net	SRERL
1	32.6	36.4	39.0	**51.7**	47.2	45.8	43.4	[49.4]
2	25.6	41.8	35.2	40.4	**44.0**	39.8	38.0	[42.1]
4	37.4	43.0	48.6	**56.0**	54.9	55.1	54.2	[55.5]
6	42.3	55.0	76.1	76.1	[77.5]	75.7	77.1	**79.4**
7	50.5	67.0	72.9	73.5	74.6	[77.2]	76.7	**78.9**
10	72.2	66.3	81.9	79.9	[84.0]	82.3	83.8	**84.5**
12	74.1	65.8	86.2	85.4	86.9	86.6	[87.2]	**88.2**
14	[65.7]	54.1	58.8	62.7	61.9	58.8	63.3	**67.3**
15	38.1	33.2	37.5	37.3	43.6	[47.6]	45.3	**50.5**
17	40.0	48.0	59.1	62.9	60.3	[62.1]	60.5	**65.1**
23	30.4	31.7	35.9	38.8	42.7	47.4	[48.1]	**50.0**
24	42.3	30.0	35.8	41.6	41.9	[55.4]	54.2	**56.5**
Avg.	45.9	48.3	55.9	58.9	60.0	[61.1]	61.0	**64.0**

表 2 DISFA 数据集上 *F*1 分数比较 (%)

AU	DRML	EAC	DSIN	JAA	ARL	LP-Net	SRERL
1	17.3	41.5	42.4	43.7	**43.9**	29.9	[43.8]
2	17.7	26.4	39.0	**46.2**	42.1	24.7	**46.2**
4	37.4	66.4	68.4	56.0	63.6	**72.7**	[67.3]
6	29.0	**50.7**	28.6	41.4	41.8	46.8	[50.1]
9	10.7	**80.5**	46.8	44.7	40.0	[49.6]	42.4
12	37.7	89.3	70.8	69.6	**76.2**	[72.9]	71.2
25	38.5	88.9	90.4	88.3	**95.2**	[93.8]	93.5
26	20.1	15.6	42.2	58.4	**66.8**	[65.0]	54.3
Avg.	26.7	48.5	53.6	56.0	**58.7**	56.9	[58.6]



(a) BP4D 数据集上 AUC 分数比较 (b) DISFA 数据集上 AUC 分数比较

图 5 BP4D 和 DISFA 数据集上 AUC 分数比较

3.2.2 面部动作单元强度预测

针对面部动作单元强度预测任务, 本文与以下当前流行的先进 AU 强度预测算法进行对比: iEAC[32], iARL[34], OR-CNN[36], CCNN-IT[37], 2DC[38], 其中 OR-CNN, CCNN-IT 和 2DC 皆属于序数模型, iMethod 代表该方法应用于面部动作单元强度预测任务, 即 iEAC 和 iARL 分别是 EAC 模型和 ARL 模型应用于面部动作单元强度预测任务时的名称. 表 3 和表 4 分别列举 FERA2015 数据集和 DISFA 数据集下的比较结果, 本文提出模型在面部动作单元强度预测任务中 (iSRERL) 取得了最优的 ICC 指标性能, 在 MAE 指标上仅次于当前先进的 iARL 算法. 通过与 iARL 算法对比, 我们可以进一步地验证 iSRERL 对于 AU 关系的推理能力: 在 MAE 指标略低于 iARL 算法的条件下, iSRERL 取得了更高的 ICC 性能, 尤其对于 DISFA 数据集中 AU5 来说, iARL 虽然具有极低的平均绝对误差 (0.04), 但却仅有 0.22 的组内相关系数; 而 iSRERL 在较高的平均绝对误差下取得了 0.44 的组内相关系数, 远优

于当前任何一种先进的方法. iSRERL 模型通过联合优化 AU 关系和特征提取, 使得模型预测结果更加接近样本真实分布, 也就具有了更高的 ICC 指标.

表 3 FERA2015 数据集结果比较

AU	ICC					MAE			
	OR-CNN	CCNN-IT	2DC	iARL	iSRERL	OR-CNN	CCNN-IT	iARL	iSRERL
6	0.60	0.75	0.76	0.72	**0.81**	1.37	1.14	0.62	**0.57**
10	0.61	0.69	0.71	0.72	**0.75**	1.39	1.30	**0.69**	**0.69**
12	0.59	0.86	0.85	0.85	**0.88**	1.37	0.99	0.51	**0.48**
14	0.25	0.40	0.45	0.44	**0.48**	1.80	1.65	**0.91**	0.94
17	0.31	0.45	0.53	0.57	**0.64**	1.19	1.08	**0.55**	0.66
Avg.	0.47	0.63	0.66	0.66	**0.71**	1.42	1.23	**0.66**	0.67

表 4 DISFA 数据集结果比较

AU	ICC					MAE			
	OR-CNN	CCNN-IT	2DC	iARL	iSRERL	OR-CNN	CCNN-IT	iARL	iSRERL
1	0.03	0.18	**0.70**	0.13	0.50	1.05	0.87	**0.30**	0.55
2	0.07	0.15	0.55	0.36	**0.64**	0.87	0.63	**0.31**	0.43
4	0.01	0.61	0.69	0.68	**0.78**	1.47	0.86	0.52	**0.47**
5	0.00	0.07	0.05	0.22	**0.44**	0.17	0.26	**0.04**	0.35
6	0.29	**0.65**	0.59	0.56	0.57	0.79	0.73	**0.36**	0.50
9	0.08	0.55	0.57	0.36	**0.61**	0.70	0.57	**0.30**	0.32
12	0.67	0.82	**0.88**	0.86	0.84	0.69	0.55	**0.31**	0.42
15	0.13	0.44	0.32	**0.52**	0.33	0.44	0.38	**0.05**	0.29
17	0.27	**0.37**	0.10	**0.37**	0.35	0.59	0.57	**0.33**	0.65
20	0.00	**0.28**	0.08	0.12	0.09	0.50	0.45	**0.08**	0.59
25	0.59	0.77	0.90	**0.96**	0.95	1.33	0.81	0.29	**0.28**
26	0.33	0.54	0.50	0.60	**0.63**	0.86	0.64	**0.26**	0.53
Avg.	0.20	0.45	0.50	0.48	**0.56**	0.79	0.61	**0.26**	0.45

另外图 6 展示了 iSRERL 与 iEAC 在 FERA2015 数据集下 MSE 指标比较, 可以看出本文提出的方法与 iEAC 有着极为接近的结果, 但在具体每个 AU 的比较中发现 iSRERL 在 AU14 的均方误差远高于 iEAC 模型, 这可能由于均方误差 MSE 这一指标容易受到异常值的干扰, 同时由于 FERA2015 数据集仅有 5 个 AU, 无法充分发挥 iSRERL 对于 AU 全局关系的推理能力.

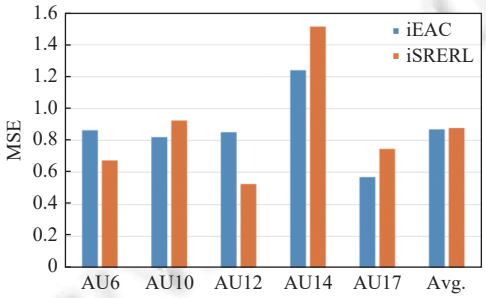


图 6 FERA2015 数据集上 MSE 结果比较

3.3 泛化性能分析

3.3.1 面部动作单元检测

BP4D 和 DISFA 数据集皆由实验室条件下采集得到的, 受到实验室环境有限性的影响 (比如在实验室中很难

引导出悲伤这种面部情绪), 数据集的样本分布与现实生活中的样本分布存在很大的差异, 因此模型的泛化性能优劣对于面部动作单元分析任务来说尤为重要, 其实是在现实场景应用时. 本节在 BP4D+数据集上验证 SRERL 在面部动作单元检测和面部动作单元强度预测两种不同任务下的泛化性能, 并与当前最先进的算法 JAA 和 ARL 进行对比. 具体地, 本节利用在 BP4D 所有数据上训练好的模型在 BP4D+数据集上进行测试.

图 7 展示了 SRERL 在 BP4D+数据集上的 $F1$ 分数, 本文提出的 SRERL 模型相比于 JAA 和 ARL 算法分别相对提高了 6.2% 和 3.8% 的 $F1$ 分数, 同时 SRERL 在大部分 AU 的 $F1$ 分数均超越这两个算法. 但是相比于 BP4D 数据集上的结果 ($F1$ 分数: 64.0%), $F1$ 分数降低了约 11.4%, 这说明本文提出的 SRERL 算法相比于当前先进的方法拥有更好的面部动作单元检测泛化性能, 但是由于数据集自身的局限性, 在跨数据集验证时候仍有一定的性能下降.

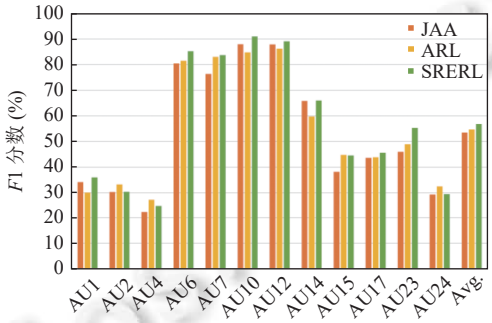


图 7 BP4D+数据集上 $F1$ 分数比较

3.3.2 面部动作单元强度预测

表 5 列举了 iSRERL 与 iARL 和 iJAA 模型在 BP4D+数据集上面部动作单元强度预测任务的泛化性能. 通过分析可以得出, 相比于 iJAA 和 iARL 模型, SRERL 取得了更好的 ICC 性能和更差的 MAE 结果, iSRERL 比 iJAA 和 iARL 模型在 ICC 上分别相对提高了 6.8% 和 3.2%. 这可能是由于 iSRERL 引入了 AU 语义关系引导特征学习, 使得预测结果与真实标签更具有一致性, 从而使得具有更高的 ICC 性能. 同时相比于 BP4D 数据集上的结果相对降低了 12.7% 的 ICC 性能和相对提高了 11.9% 的 MAE (误差指标, 越低越好) 结果, 综上可以认为本文提出的 iSRERL 模型虽然取得了与当前最先进算法相匹配的泛化性能, 但在跨数据集验证时候仍有不少的性能损失.

表 5 BP4D+数据集上 AU 强度预测实验结果

AU	ICC			MAE		
	iJAA	iARL	iSRERL	iJAA	iARL	iSRERL
6	0.72	0.78	**0.79**	0.63	**0.58**	0.68
10	0.79	0.77	**0.83**	0.59	**0.55**	0.66
12	0.82	0.82	**0.86**	**0.63**	0.70	0.68
14	0.14	**0.19**	0.10	0.79	**0.71**	1.16
17	0.45	0.50	**0.56**	0.38	**0.33**	0.56
Avg.	0.59	0.61	**0.63**	0.60	**0.57**	0.75

3.4 遮挡性能测试

在现实场景中, 人脸经常容易被其他物体所遮挡, 比如眼睛, 手, 口罩等, 而由于面部信息丢失, 遮挡问题对于面部动作单元分析来说仍是严峻挑战. 为了验证 SRERL 在遮挡条件下的性能, 本节将输入图像部分遮挡, 并利用在未遮挡条件下训练好的模型进行测试. 如图 8 所示, 通过遮挡上下左右 4 部分的脸模拟现实场景中对人脸的遮挡, 同时为了验证有效性, 本文还加入了全部遮挡的情况作为输入异常输入. 图 9 分别展示了本文提出模型与 EAC[32]和 ARL[34]在 5 种不同遮挡条件下比较的结果.

通过观察图 9 可发现, 在上下左右人脸遮挡实验中, 我们的方法 SRERL 均取得了最优的 $F1$ 得分结果. 证明了 SRERL 算法的特征传播模块有效的利用未遮挡区域的 AU 情况和不同 AU 间的语义关系成功引导遮挡区域 AU 的预测, 进而提升了算法的鲁棒性. 在对人脸图像全部遮挡进行异常输入测试实验时, 由于输入图像不存在任何面部信息, 因此理论上预测结果应符合随机猜测. 图 9(e) 展示了与 EAC 和 ARL 方法的对比, 可以看到 3 种算法的 $F1$ 分数都很低, 本文提出的方法在 $F1$ 分数上高于 ARL 而低于 EAC, 进一步验证了实验的合理性.

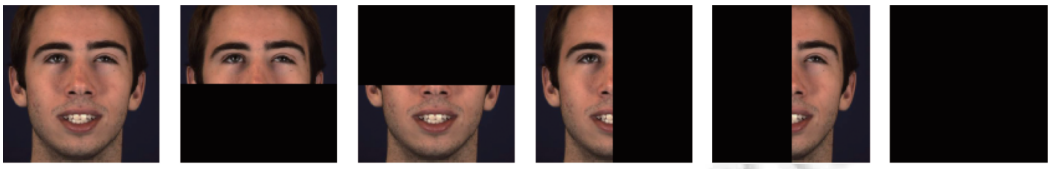
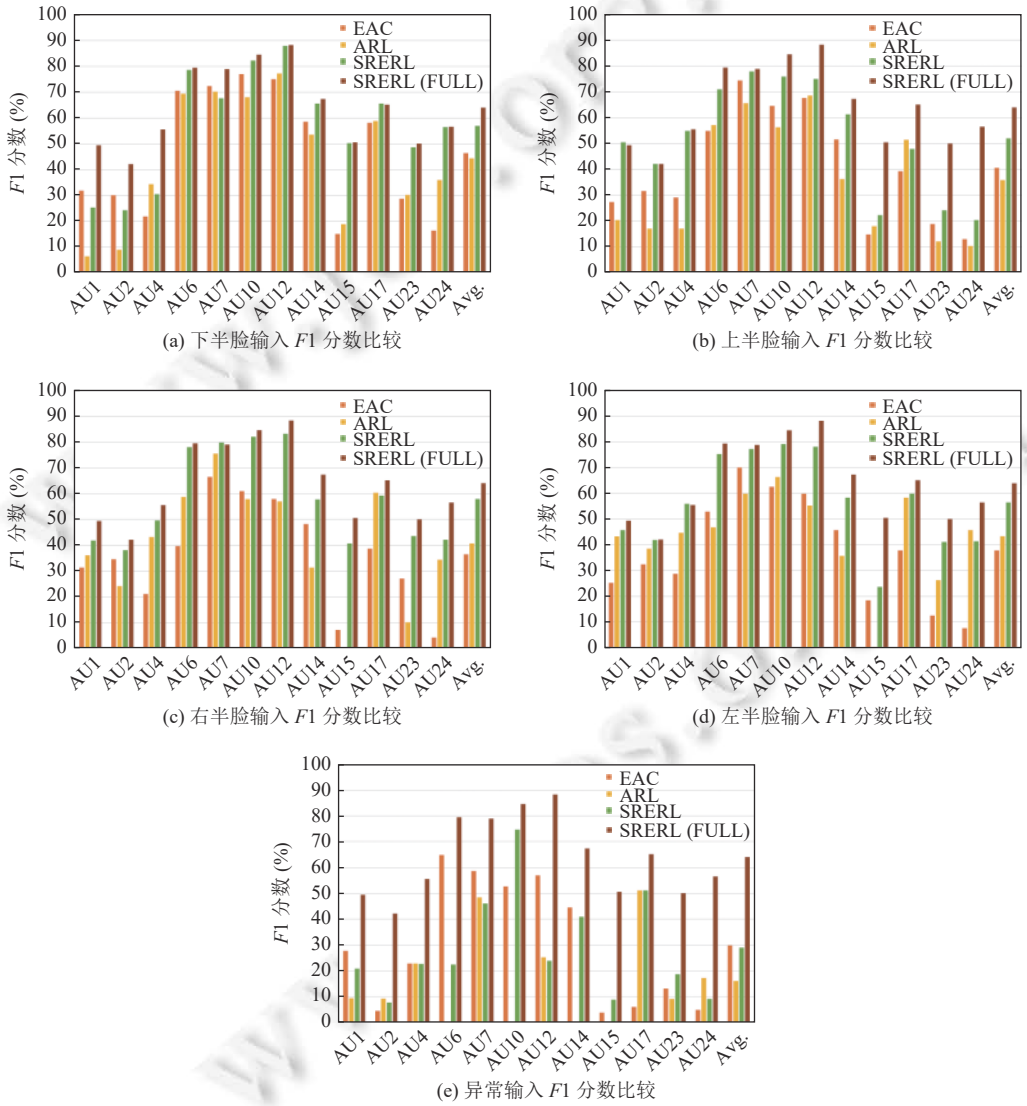


图 8 面部遮挡示意图



(a) 下半脸输入 $F1$ 分数比较

(b) 上半脸输入 $F1$ 分数比较

(c) 右半脸输入 $F1$ 分数比较

(d) 左半脸输入 $F1$ 分数比较

(e) 异常输入 $F1$ 分数比较

图 9 遮挡测试下 $F1$ 分数比较

3.5 消融实验

为了验证 SRERL 每个模块的合理性和有效性, 本节通过消融实验详细分析每个组件的作用. 本节将 SRERL 拆分成加权损失函数、区域学习、多尺度特征融合和语义关系传播共 4 个组件, 并由这 4 个组件组合成 VGG, VGG_BL, SS_RL, MS_RL 和 SRERL 这 5 种方法. 本节在面部动作单元检测和面部动作单元强度预测两个不同任务上逐个分析每个组件的用处, 其中 iMethod 代表该方法用于面部动作单元强度预测任务.

3.5.1 加权损失函数的有效性

样本正负比例不平衡是面部动作单元分析任务中的常见问题, 本文尝试采用加权损失函数解决数据不平衡问题, 表 6 和图 10 分别展示了 BP4D 数据集上 AU 检测消融实验结果和 FERA2015 数据集上 AU 强度预测消融实验结果. 如表 6 所示, 采用加权损失函数的方法 VGG_BL 相比于基线方法 VGG 在面部动作单元检测任务中分别相对提高了 3.6% 的 $F1$ 分数和 0.3% 的 AUC 指标. 通过具体分析可以发现, 大部分正样本比例较少的 AU 有明显的性能提升, 例如 VGG_BL 在 AU24 上分别相对提高了 13.2% 的 $F1$ 分数和 2.1% 的 AUC 结果, 在 AU23 上分别提高了 21.8% 和 2.2% 等. 这说明本文提出的加权损失函数能在一定程度上有效改善面部动作单元检测任务中数据不平衡问题.

表 6 BP4D 数据集上 AU 检测消融实验结果

AU	$F1$分数 (%)					AUC (%)				
	VGG	VGG_BL	SS_RL	MS_RL	SRERL	VGG	VGG_BL	SS_RL	MS_RL	SRERL
1	40.7	41.7	47.6	47.4	49.4	74.8	70.1	77.4	79.0	77.4
2	32.9	36.0	38.0	42.4	42.1	70.7	67.5	73.5	74.7	74.5
4	45.8	49.7	55.1	54.7	55.5	77.2	78.0	83.6	83.6	84.3
6	78.6	78.2	77.8	78.6	79.4	88.5	88.5	88.7	88.5	89.3
7	76.5	76.1	76.5	78.3	78.9	82.5	82.7	85.4	85.5	86.2
10	84.6	82.5	84.9	84.2	84.5	86.2	86.3	88.5	88.6	88.6
12	88.0	85.9	87.8	86.6	88.2	94.0	93.6	94.1	94.4	94.6
14	63.0	63.4	67.3	69.0	67.3	66.1	70.2	73.7	74.6	76.0
15	40.4	46.7	45.3	47.7	50.5	78.3	79.7	81.2	81.0	81.1
17	59.9	61.7	65.1	61.7	65.1	77.2	76.8	80.2	79.6	80.3
23	34.8	42.4	47.2	48.2	50.0	73.2	74.8	79.0	79.6	79.6
24	46.8	53.0	54.3	55.8	56.5	85.8	87.6	88.5	89.2	89.3
Avg.	57.7	59.8	62.4	62.8	64.0	79.5	79.7	82.8	83.2	83.4

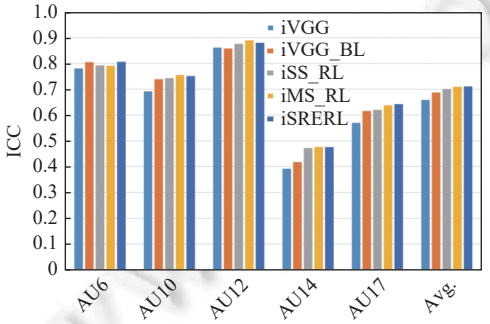


图 10 FERA2015 数据集上 AU 强度预测消融实验 ICC 结果

为了验证加权损失函数在面部动作单元强度预测任务中的有效性, 本文在 FERA2015 数据集上进行详细分析. 如图 10 所示, 除了样本比较较为均衡的 AU12, iVGG_BL 在其余 4 个 AU 上皆取得较好的性能提升, 总体 ICC 相比 iVGG 提高了 4.3%, 这进一步证明了本文设计的加权损失函数在面部动作单元强度预测任务中的通用性.

3.5.2 区域学习模块的有效性

如表 6 所示, 融入区域学习模块的 SS_RL 模型相比于参数共享的 VGG_BL 模型在大部分 AU 上具有更高的面部动作单元检测 $F1$ 分数和 AUC 性能, 尤其是对于一些表观特征在全图中不太明显的 AU, 例如 AU4 (压低眉毛), AU12 (拉升嘴角) 和 AU1 (提升内眉梢) 等. 以 AU4 为例, SS_RL 比 VGG_BL 在 $F1$ 和 AUC 上分别相对提升了 10.9% 和 7.2%, 这足以说明基于区域学习模块的自适应特征学习在面部动作单元检测任务中的有效性. 同样在图 10 中, iSS_RL 相比 iVGG_BL 提高了 2.0% 的 ICC 性能.

3.5.3 多尺度特征融合的有效性

为了验证多尺度特征融合的有效性, 本文将基于多尺度的 MS_RL 和单一尺度的 SS_RL 模型进行对比. 表 6 表明 MS_RL 比 SS_RL 总体提高了 0.6% 的 $F1$ 分数和 0.48% 的 AUC 性能, 尤其是对于占用人脸面积较小的 AU1, AU2, AU12 和 AU14 等, 其中 MS_RL 相比于 SS_RL 在 AU1 和 AU2 上分别提升了 2.1% 和 1.6% 的 AUC 性能, 这符合多尺度特征融合能提升小目标检测结果的猜想. 更进一步地, 为了说明本文特征提取模块的作用, 本文在 DISFA 数据集上进行面部动作单元检测和面部动作单元强度预测两种任务下的消融实验. 通过分析图 11 和表 7 可以得出, 结合多尺度特征和区域学习通道的特征提取模块 MS_RL (iMS_RL) 相较于基础的加权传统网络 VGG_BL (iVGG_BL) 在两种不同任务下分别提高了 9.9% 的 $F1$ 和 7.4% 的 ICC.

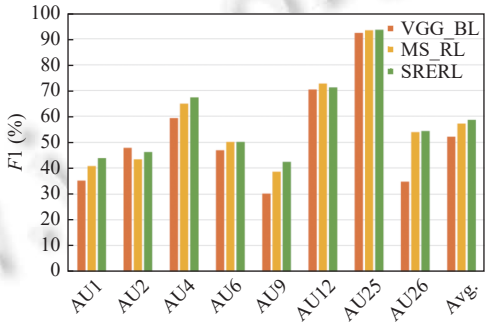


图 11 DISFA 数据集上 AU 检测消融实验 $F1$ 结果

表 7 DISFA 数据集上 AU 强度预测消融实验结果

AU	ICC			MAE			MSE		
	iVGG_BL	iMS_RL	iSRERL	iVGG_BL	iMS_RL	iSRERL	iVGG_BL	iMS_RL	iSRERL
1	0.635	0.492	0.505	0.602	0.678	0.553	0.653	1.114	0.895
2	0.641	0.682	0.641	0.616	0.279	0.430	0.593	0.285	0.441
4	0.685	0.786	0.777	0.595	0.391	0.474	0.611	0.429	0.506
5	0.349	0.445	0.443	0.656	0.326	0.348	0.558	0.208	0.239
6	0.518	0.525	0.571	0.449	0.661	0.502	0.509	0.795	0.543
9	0.507	0.565	0.609	0.325	0.400	0.319	0.229	0.361	0.268
12	0.840	0.825	0.842	0.368	0.505	0.421	0.299	0.472	0.365
15	0.213	0.289	0.326	0.378	0.239	0.292	0.257	0.170	0.200
17	0.272	0.354	0.349	0.725	0.574	0.647	0.817	0.691	0.841
20	0.075	0.091	0.087	0.501	0.605	0.586	0.377	0.776	0.777
25	0.935	0.945	0.949	0.409	0.295	0.285	0.266	0.182	0.177
26	0.501	0.629	0.626	0.678	0.467	0.533	0.703	0.434	0.575
Avg.	0.514	0.552	0.560	0.525	0.452	0.449	0.490	0.493	0.486

3.5.4 语义关系融合的有效性

为了验证语义关系模块的有效性, 本文基于 AU 语义关系建模的 SRERL 模型与无关系建模的 MS_RL 进行对比. 如表 6 和图 11 所示, SRERL 相比 MS_RL 在面部动作单元检测任务上分别相对提高了 1.9% 和 1.6% 的 $F1$

分数. 同时针对 AU 个体检测性能的分析可以看出, AU 语义关系引导对于提高检测精度起着至关重要的作用, 例如 AU6 和 AU12 有着很强的共生性, 在 BP4D 数据集中, AU6 和 AU12 的 $F1$ 分数分别提高了 1.0% 和 1.8%, AUC 性能分别提高了 0.9% 和 0.2%. 另一方面, AU4 和 AU12 具有很强的互斥性, AU4 在 BP4D 数据集和 DISFA 数据集上分别提高了 1.5% 和 3.6% 的 $F1$ 分数与 0.8% 和 1.0% 的 AUC.

AU 语义关系同样能提高面部动作单元强度预测性能, 如表 7 所示, 基于 AU6 和 AU12 的共生性, AU6 和 AU12 分别在 DISFA 上提升了 8.8% 和 2.1% 的 ICC 性能, 降低了 24% 和 16.6% 的 MAE; 而基于 AU4 和 AU9 的互斥性, 相比于 iMS_RL 模型, AU9 在 FERA2015 数据集和 DISFA 数据集上都提升了 7.8% 的 ICC 性能. 总体而言, iSRERL 模型比 iMS_RL 具有更高的 ICC 指标和更低的 MAE 指标.

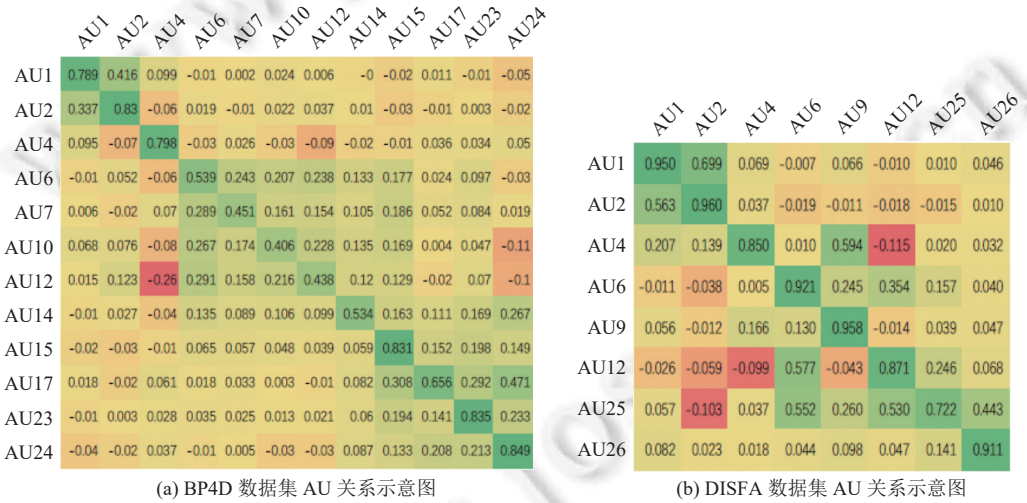
综合上述实验结果可以很好地表明, 本文提出的基于 AU 语义关系的模块能很好地利用 AU 之间的全局关系极大地增强人脸区域特征表示, 使得模型学到的特征更具有区分性从而取得更高的精度.

3.6 可解释性分析

为了进一步地分析模型合理有效性, 本节将对模型的可解释性进行详细的分析, 具体分为面部动作单元关系可视化和注意力机制可视化两个部分.

3.6.1 面部动作单元关系可视化

图 12 和图 13 分别可视化了两种不同面部动作单元分析任务下的 AU 关系, 其中绿色越深代表共生关系越强, 红色越深代表互斥关系越强. 由图 12(a) 可以看出其中具有较强共生关系的由 AU1 (提升内眉梢) 和 AU2 (提升外眉梢)、AU6 (提高脸颊) 和 AU12 (拉升嘴角)、AU17 (提升下颏突) 和 AU24 (挤压嘴唇) 等, 而 AU4 (压低眉毛) 和 AU12 (拉升嘴角) 与 AU4 (压低眉毛) 和 AU2 (提升外眉梢) 等具有较强的互斥关系, 这符合面部动作编码系统中人脸肌肉结构性的设置. 然而由于数据集多样性的限制, 在图 12(a) 和图 13(a) 中皆未能统计出 AU12 (拉伸嘴角) 和 AU15 (降低嘴角) 的互斥性, 这一情况也与当前研究结果[12]类似.

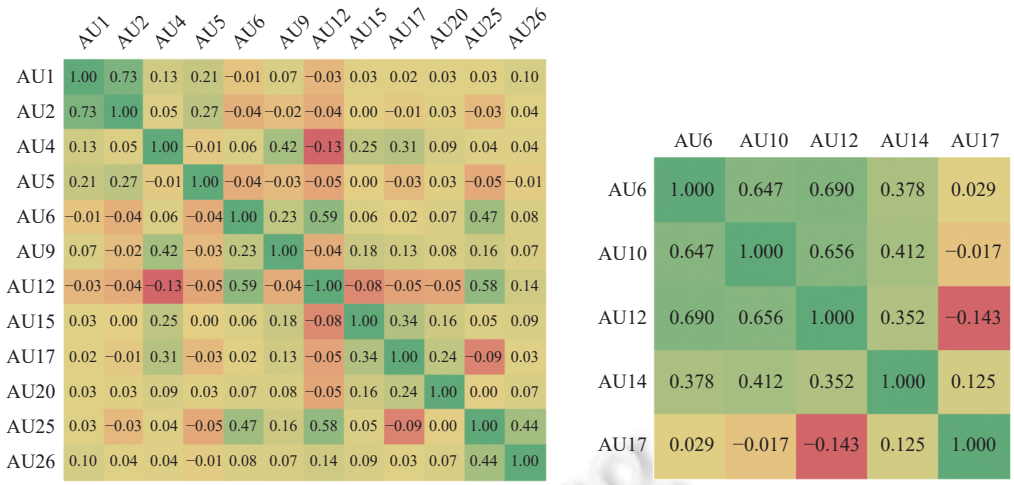


(a) BP4D 数据集 AU 关系示意图 (b) DISFA 数据集 AU 关系示意图

图 12 面部动作单元检测条件下 AU 关系示意图

3.6.2 注意力机制可视化

为了进一步验证模型的可解释性, 本文尝试对区域学习通道的注意力机制进行可视化, 并以热力图的形式展示每个面部动作单元所对应的激活区域. 具体地, 本文单独训练了 MS_RL 模型, 在区域学习模块中本文从 14×14 大小的全局特征图上提取了 6×6 大小的局部特征图, 每个局部特征图对应全图 18% 大小, 然后每一个区域通道都接以单独的损失函数进行训练, 最后采用 Grad-CAM[39]可视化每个 AU 对应的类激活图. 图 14 展示了 12 个 AU 在不同肤色和性别输入下的类激活图, 其对应的激活区域都较好的匹配相应的 AU 中心位置.



(a) DISFA 数据集 AU 关系示意图　　(b) FERA2015 数据集 AU 关系示意图

图 13　面部动作单元强度条件下 AU 关系示意图

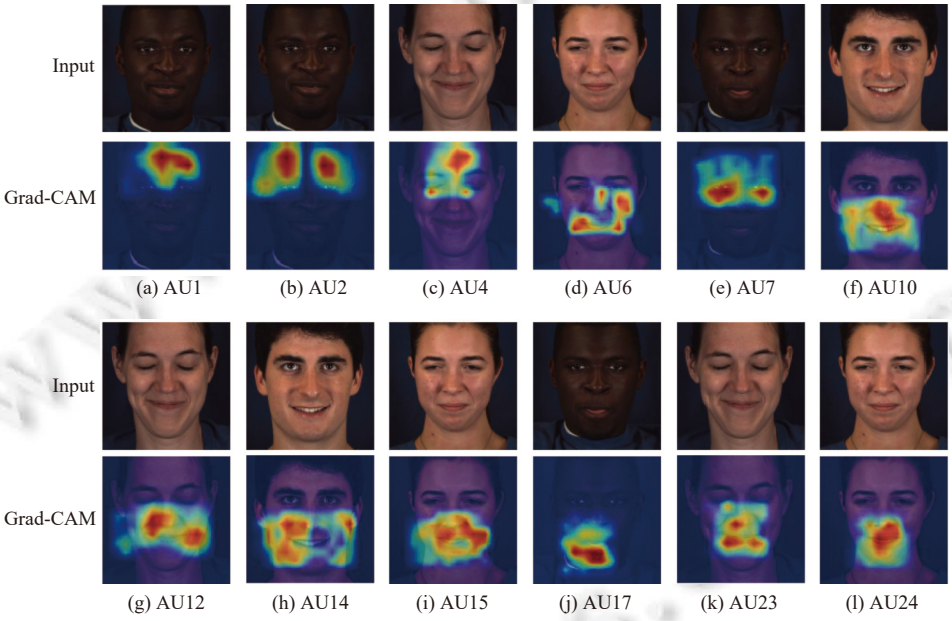


图 14　注意力机制可视化示意图

4　总结与展望

本文针对面部动作单元分析问题, 通过分析不同面部行为中 AU 之间的共生互斥性对其进行关系建模, 提出了基于语义关系的表征学习算法 SRERL, 包括多尺度的主干神经网络模块, 基于注意力机制的区域学习模块和基于语义关系的图神经网络模块. 本文在多个公开的面部动作单元数据集上进行 AU 检测和 AU 强度预测两个任务的实验并与现有先进算法进行对比, 充分论证了模型的有效性. 更进一步, 本文还通过一系列实验证明模型的泛化性, 鲁棒性和各个模块的合理有效性.

尽管本文针对面部动作单元分析问题在公开数据集上取得了相较以往算法更高的性能, 但事实上面部动作单元分析领域还远没有人脸表情识别成熟, 在准确率上也远没有达到可以落地的效果. 综合考虑现有面部动作单元

分析数据集和算法的优缺点, 本文认为未来面部动作单元分析问题可以在以下几方面展开研究.

1) 现有面部动作单元数据集都是基于实验条件下录制的, 受实验室环境的局限性, 这些数据集的样本分布与现实场景中的样本分布存在很大差异. 并且大部分数据集样本和标注单一, 而不同数据集之间标注 AU 的种类又各不一样, 这极大地限制了模型的泛化性能和跨数据集交叉训练. 因此收集并标注一个基于现实场景下的大型面部动作单元数据集对于面部动作单元分析问题来说至关重要.

2) 现有面部动作单元分析算法大都基于单帧图像, 而视频相比图像拥有更丰富的运动信息, 同时人脸长相的差异性是面部动作单元分析任务中的巨大挑战, 如何有效利用视频中丰富的运动信息并消除人脸长相的差异性带来的影响是提高面部动作单元分析模型性能的关键.

3) 虽然面部动作单元的标注及其困难, 但是带有表情标注的人脸图像却可以轻易获得, 如何有效利用这些没有面部动作单元信息标注的数据以提高面部动作单元分析准确性是当前主流的研究方向之一; 另一方面如何将现有面部动作单元分析算法结合多模态的信息输入, 并应用于微表情分析等具体场景是现有面部动作单元分析问题落地的关键.

References:

[1] Ekman P, Friesen WV. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press, 1978.

[2] Valstar M, Pantic M. Fully automatic facial action unit detection and temporal analysis. In: Proc. of the 2006 Conf. on Computer Vision and Pattern Recognition Workshop. New York: IEEE, 2006. 149. [doi: 10.1109/CVPRW.2006.85]

[3] Zhao KL, Chu WS, De La Torre F, Cohn JF, Zhang HG. Joint patch and multi-label learning for facial action unit detection. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 2207–2216. [doi: 10.1109/CVPR.2015.7298833]

[4] Jiang BH, Valstar MF, Pantic M. Action unit detection using sparse appearance descriptors in space-time video volumes. In: Proc. of the 2011 IEEE Int'l Conf. on Automatic Face & Gesture Recognition. Santa Barbara: IEEE, 2011. 314–321. [doi: 10.1109/FG.2011.5771416]

[5] Zhao KL, Chu WS, Zhang HG. Deep region and multi-label learning for facial action unit detection. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3391–3399. [doi: 10.1109/CVPR.2016.369]

[6] Li W, Abtahi F, Zhu ZG. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6766–6775. [doi: 10.1109/CVPR.2017.716]

[7] Bishay M, Patras I. Fusing multilabel deep networks for facial action unit detection. In: Proc. of the 12th IEEE Int'l Conf. on Automatic Face & Gesture Recognition. Washington: IEEE, 2017. 681–688. [doi: 10.1109/FG.2017.86]

[8] Chu WS, De La Torre F, Cohn JF. Modeling spatial and temporal cues for multi-label facial action unit detection. arXiv:1608.00911, 2016.

[9] Tong Y, Liao WH, Ji Q. Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007, 29(10): 1683–1699. [doi: 10.1109/TPAMI.2007.1094]

[10] Wang ZH, Li YQ, Wang SF, Ji Q. Capturing global semantic relationships for facial action unit recognition. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 3304–3311. [doi: 10.1109/ICCV.2013.410]

[11] Li YJ, Tarlow D, Brockschmidt M, Zemel RS. Gated graph sequence neural networks. In: Proc. of the 4th Int'l Conf. on Learning Representations. San Juan: ICLR, 2015.

[12] Corneanu C, Madadi M, Escalera S. Deep structure inference network for facial action unit recognition. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 309–324. [doi: 10.1007/978-3-030-01258-8_19]

[13] Du SC, Tao Y, Martinez AM. Compound facial expressions of emotion. Proc. of the National Academy of Sciences of the United States of America, 2014, 111(15): E1454–E1462.

[14] Peng GZ, Wang SF. Weakly supervised facial action unit recognition through adversarial training. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2188–2196. [doi: 10.1109/CVPR.2018.00233]

[15] Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. In: Proc. of the 2015 British Machine Vision Conf. BMVA Press, 2015. 1–12.

[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2014.

[17] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the 2017 IEEE Conf. on

Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269. [doi: 10.1109/CVPR.2017.243]

[18] Lin TY, Dollár P, Girshick R, He KM, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944. [doi: 10.1109/CVPR.2017.106]

[19] Chen TS, Lin L, Chen RQ, Wu Y, Luo XN. Knowledge-embedded representation learning for fine-grained image recognition. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI, 2018.

[20] Wang ZX, Chen TS, Ren JSJ, Yu WH, Cheng H, Lin L. Deep reasoning with knowledge graph for social relationship understanding. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI, 2018. 1021–1028.

[21] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555, 2014.

[22] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[23] Zhang X, Yin LJ, Cohn JF, Canavan S, Reale M, Horowitz A, Liu P, Girard JM. BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database. Image and Vision Computing, 2014, 32(10): 692–706. [doi: 10.1016/j.imavis.2014.06.002]

[24] Mavadati SM, Mahoor MH, Bartlett K, Trinh P, Cohn JF. DISFA: A spontaneous facial action intensity database. IEEE Trans. on Affective Computing, 2013, 4(2): 151–160. [doi: 10.1109/T-AFFC.2013.4]

[25] Valstar MF, Almaev T, Girard JM, McKeown G, Mehu M, Yin LJ, Pantic M, Cohn JF. FERA 2015-second facial expression recognition and analysis challenge. In: Proc. of the 2015 IEEE Int'l Conf. and Workshops on Automatic Face and Gesture Recognition. Ljubljana: IEEE, 2015. 1–8. [doi: 10.1109/FG.2015.7284874]

[26] Zhang Z, Girard JM, Wu Y, Zhang X, Liu P, Ciftci U, Canavan S, Reale M, Horowitz A, Yang HY, Cohn JF, Ji Q, Yin LJ. Multimodal spontaneous emotion corpus for human behavior analysis. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3438–3446. [doi: 10.1109/CVPR.2016.374]

[27] Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 1979, 86(2): 420–428. [doi: 10.1037/0033-2909.86.2.420]

[28] Bradski G, Kaehler A. Learning OpenCV: Computer Vision with the OpenCV Library. Sebastopol: O'Reilly Media, 2008.

[29] King DE. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 2009, 10: 1755–1758.

[30] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin ZM, Gimelshein N, Antiga L. PyTorch: An imperative style, high-performance deep learning library. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 721.

[31] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2014.

[32] Li W, Abtahi F, Zhu ZG, Yin LJ. EAC-Net: Deep nets with enhancing and cropping for facial action unit detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2583–2596. [doi: 10.1109/TPAMI.2018.2791608]

[33] Shao ZW, Liu ZL, Cai JF, Ma LZ. Deep adaptive attention for joint facial action unit detection and face alignment. In: Proc. of the European Conf. on Computer Vision. Munich: Springer, 2018. 725–740. [doi: 10.1007/978-3-030-01261-8_43]

[34] Shao ZW, Liu ZL, Cai JF, Wu YS, Ma LZ. Facial action unit detection using attention and relation learning. IEEE Trans. on Affective Computing, 2022, 13(3): 1274–1289. [doi: 10.1109/TAFFC.2019.2948635]

[35] Niu XS, Han H, Yang SF, Huang Y, Shan SG. Local relationship learning with person-specific shape regularization for facial action unit detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11909–11918. [doi: 10.1109/CVPR.2019.01219]

[36] Niu ZX, Zhou M, Wang L, Gao XB, Hua G. Ordinal regression with multiple output CNN for age estimation. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4920–4928. [doi: 10.1109/CVPR.2016.532]

[37] Walecki R, Rudovic O, Pavlovic V, Schuller B, Pantic M. Deep structured learning for facial action unit intensity estimation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5709–5718. [doi: 10.1109/CVPR.2017.605]

[38] Tran DL, Walecki R, Rudovic O, Eleftheriadis S, Schuller B, Pantic M. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 3209–3218. [doi: 10.1109/ICCV.2017.346]

[39] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

李冠彬(1986－), 男, 博士, 副教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 机器学习.

朱鑫(1995－), 男, 硕士生, 主要研究领域为计算机视觉.

张锐斐(1998－), 男, 硕士生, 主要研究领域为计算机视觉.

林倞(1981－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为计算机视觉, 机器学习.