计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(8):42−53 [doi: 10.15888/j.cnki.csa.009183]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 MIFNet 的婴儿面部表情识别①

耿　磊[1,3], 齐婷婷[2,3], 张　芳[1,3], 肖志涛[1,3], 李月龙[4]

[1](天津工业大学 生命科学学院, 天津 300387)
[2](天津工业大学 电子与信息工程学院, 天津 300387)
[3](天津市光电检测技术与系统重点实验室, 天津 300387)
[4](天津工业大学 计算机科学与技术学院, 天津 300387)
通信作者: 张　芳, E-mail: hhzhangfang@126.com

摘　要: 婴儿面部表情智能化识别, 可辅助看护人员更好地关注婴儿的身心健康. 由于婴儿面部线条流畅且五官锐感偏弱导致面部表情类间相似性高于成人, 为了解决类间相似性高的问题, 提出多尺度信息融合网络. 该网络整体分为 2 个阶段: 在第 1 阶段使用融合模块在空间域与通道域双重维度下融合局部特征与全局特征, 增强特征的表达能力; 在第 2 阶段采用自适应深度中心损失, 利用注意力机制估计融合特征的权重用以指导中心损失, 促进婴儿表情特征的类内紧凑和类间分离. 实验结果表明, 多尺度信息融合网络在婴儿面部表情数据集中识别准确率达到 95.46%, 在 AUC、召回率和 $F1$ 得分 3 个评价指标上分别达到 99.07%、95.88% 和 95.89%, 与现有面部表情识别网络相比, 识别效果最优. 将多尺度信息融合网络在公开面部表情数据集上进行泛化性实验, 准确率达到 89.87%.

关键词: 婴儿; 表情识别; 自适应深度中心损失; 多尺度特征融合; 注意力机制

引用格式: 耿磊,齐婷婷,张芳,肖志涛,李月龙.基于 MIFNet 的婴儿面部表情识别.计算机系统应用,2023,32(8):42–53. http://www.c-s-a.org.cn/1003-3254/9183.html

Facial Expression Recognition of Infants Based on MIFNet

GENG Lei[1,3], QI Ting-Ting[2,3], ZHANG Fang[1,3], XIAO Zhi-Tao[1,3], LI Yue-Long[4]

[1](School of Life Sciences, Tiangong University, Tianjin 300387, China)
[2](School of Electronics and Information Engineering, Tiangong University, Tianjin 300387, China)
[3](Tianjin Key Laboratory of Optoelectronic Detection Technology and Systems, Tianjin 300387, China)
[4](School of Computer Science and Technology, Tiangong University, Tianjin 300387, China)

Abstract: The intelligent recognition of infant facial expressions can help caregivers to better pay attention to the physical and mental health of infants. Due to the smooth facial lines and weak sharpness of facial features, the inter-class similarity of infants' facial expressions is higher than that of adults. To address the problem of high inter-class similarity, this study proposes a multi-scale information fusion network. The network is divided into two stages as a whole. In the first stage, the fusion module is applied to fuse local features with global features in the dual dimensions of both spatial and channel domains to enhance the expression ability of features. In the second stage, the self-adaptive deep centre loss is employed to estimate the weights of fused features based on the attentional mechanism, thus guiding the center loss and promoting the intra-class compactness and inter-class separation of infant expression features. The experimental results show that the multi-scale information fusion network achieves a recognition accuracy of 95.46% in the infant facial expressions dataset, reaching 99.07%, 95.88%, and 95.89% in the three evaluation metrics of AUC, *recall*, and $F1$ score respectively. The recognition effectiveness is optimal compared with the existing facial expression recognition networks. The generalization experiments of the multi-scale information fusion network are conducted on the public facial expressions dataset, with an

① 基金项目: 国家自然科学基金 (61771340)
收稿时间: 2023-01-16; 修改时间: 2023-02-13; 采用时间: 2023-03-03; csa 在线出版时间: 2023-06-09
CNKI 网络首发时间: 2023-06-12

accuracy of 89.87%.

Key words: infants; facial expression recognition; self-adaptive deep centre loss; multiscale feature fusion; attentional mechanism

伴随着人工智能逐渐渗透到各种领域, 以婴儿为受益主体的相关产业链逐渐变得智能化. 婴儿的语言中枢神经系统尚未发育成熟, 主要通过面部表情和肢体动作来表达意图. 同时婴儿尚未具备应对突发情况的能力, 在面对危险和身体不适时, 都无法进行相应的处理. 因此在婴儿智能看护系统中增加表情识别这一功能, 不仅可以帮助看护人员关注婴儿的状态, 还可以辅助看护人员对婴儿的突发情况进行及时有效地处理, 有助于看护人员在日常生活中给予婴儿充分地呵护, 对于婴儿的培育和健康状态的分析具有重要意义.

婴儿的面部表情识别越来越受到研究人员的关注, 并取得一定成果. Fang 等人[1] 结合图像处理和语音处理获取的信息, 提出婴儿情绪识别系统, 但识别功能具有一定的局限性. Zhang 等人[2] 提出一种用于婴儿面部表情识别的多分支融合网络, 利用改进的 LBP 和 Sobel 边缘检测算子分别获取图像水平方向和垂直方向的 3 个特征图, 将其输入到多分支网络模型获得婴儿表情类别. Zamzami 等人[3] 对婴儿的面部变形方向进行研究分析, 将面部变形作为分类的主要特征用于训练 KNN 和 SVM 分类器, 最终将婴儿的表情分为痛苦和无痛苦两个类别. 目前关于婴儿的研究主要是临床方向的疼痛表情分析, 医护人员根据婴儿疼痛表情的指标判断婴儿的状态, 并采取相应措施进行处理, 避免长期的疼痛刺激对婴儿大脑发育造成无法逆转的伤害[4]. 这些方法并不能有效地识别婴儿在日常生活状态下的面部表情, 本文是以婴儿的面部表情识别应用于日常看护设备为出发点进行研究, 辅助看护人员更好地关注婴儿状态.

现有的日常状态下面部表情识别的深度学习方法主要是以成人为主体, 婴儿与成人之间存在显著的个体差异. 成人的面部轮廓清晰、五官立体且敏锐; 婴儿面部光滑、线条流畅且五官锐感较弱, 导致婴儿不同表情类间相似性高. 同时婴儿面部姿态存在多变性且现实场景复杂多样. 上述问题均会影响婴儿面部表情识别的准确率. 为了降低上述因素的影响, 本文提出一种基于多尺度信息融合网络 (multi-scale information fusion network, MIFNet) 的算法, 该算法主要创新点如下.

1) 设计卷积神经网络分支与自注意力网络分支并行提取婴儿面部表情的局部特征与全局特征, 增强婴儿面部整体性特征的表达, 同时解决婴儿面部姿态多样性造成的特征提取不全面的问题.

2) 设计空间全局注意力模块提升婴儿面部关键区域的特征表达, 在空间域对特征进行融合用于后续分类; 采用通道注意力模块学习婴儿表情特征每个通道的重要程度和各通道之间的联系, 并为每个通道赋予权重系数, 在通道域对特征进行融合用于后续分类. 双重维度下对特征进行融合降低场景多样化的影响.

3) 针对婴儿表情类间相似性高的问题, 设计自适应深度中心损失, 在嵌入空间中缩短同一类别婴儿表情特征之间的距离并拓宽不同类别之间的距离, 实现婴儿表情特征类内紧凑与类间分离. 联合 Softmax 损失进行优化判定婴儿表情的最终类别.

1 相关工作

1.1 面部表情特征的提取

卷积神经网络 (convolutional neural network, CNN) 在面部表情识别任务中具有鲁棒性, 因此其得到广泛的应用[5]. CNN 结构擅长捕捉卷积感受野范围内的局部特征, 想要获得图像中特征的全局依赖关系必须增加卷积层的深度[6]. 理论上来说通过增加网络的深度 CNN 感受野可以覆盖整张图像, 而堆叠过深的卷积层会增加模型参数量引发过拟合问题[7]. Transformer 是由 Vaswani 等人[8] 提出的用于自然语言处理任务的网络模型. 该模型直接跳出卷积网络的思维, 仅靠自注意力机制搭建网络结构, 获取图像的全局依赖性. 其中 Swin Transformer[9] 将特征图划分成多个窗口区域, 将自注意力机制的计算限制在不重叠的局部窗口, 同时使用移位窗口方案允许跨窗口连接维持局部特征之间的联系, 从而获得图像中特征的全局信息.

1.2 面部表情特征的融合

在神经网络中引入注意力机制融合面部表情特征可以在一定程度上提高面部表情识别的准确性, 在复

杂多样的场景下, 注意力机制可以使网络更加关注面部的显著特征[10]. Wang 等人[11] 提出一个区域注意网络 (region attention network, RAN), 利用关系注意力模块将局部信息和全局背景相关联, 有效地解决面部表情识别任务中的遮挡和姿势变化问题. Gera 等人[12] 提出一种端到端的空间通道注意力网络来融合表情特征中空间与通道的信息, 提高面部表情识别的准确性. Wang 等人[13] 提出一个自愈网络 (self-cure network, SCN) 用来抑制面部表情的不确定性, 该网络利用注意力机制对每个训练的样本进行加权, 防止网络过度拟合标记不正确的样本.

1.3 面部表情识别的损失

在面部表情识别算法中引入深度度量学习的方法, 可以增强 Softmax 损失的辨别能力[14], 提高面部表情识别的准确率. 中心损失是一种被广泛采用的深度度量学习方法[15], 中心损失的目标函数最小化深度特征与其相应类中心之间的簇内平方和 (within cluster sum of squares, WCSS), 将嵌入空间划分为 K 个簇以解决 K 分类问题. 一个由 m 个样本组成的小批量训练样本, 设 $x_i=[x_{i1}, x_{i2}, \cdots, x_{id}]^{\mathrm{T}}\in\mathbb{R}^d$ 是 y_i 的第 i 个样本的深度特征向量, 其中, $y_i\in\{1,\cdots,K\}$ 和 $c_{yi}=[c_{yi1}, c_{yi2}, \cdots, c_{yid}]^{\mathrm{T}}\in\mathbb{R}^d$ 是其对应的类中心, 中心损失最小化的标准定义如式 (1) 所示.

$$L_{\mathrm{C}}=\frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{d}\|x_{ij}-c_{y_{ij}}\|_2^2 \tag{1}$$

Cai 等人[16] 提出的 Island 损失, 在中心损失的基础上增加一个额外的目标函数, 增强网络对面部表情深度特征的学习能力. 受此启发, Li 等人[17] 提出的 Separate 损失是中心损失和 Island 损失的余弦版本. Separate 损失中的类内损失和类间损失将属于同一面部表情类别的特征之间的余弦相似度最大化, 并在嵌入空间中最小化面部表情类中心之间的余弦相似度.

2 本文方法

为了辅助看护人员在日常生活中更好地关注婴儿状态, 本文设计了多尺度信息融合网络 MIFNet 用于婴儿面部表情识别. MIFNet 网络整体结构主要分为特征提取模块和由自适应深度中心损失与 Softmax 损失组成的联合损失两部分, 其框架如图 1 所示.

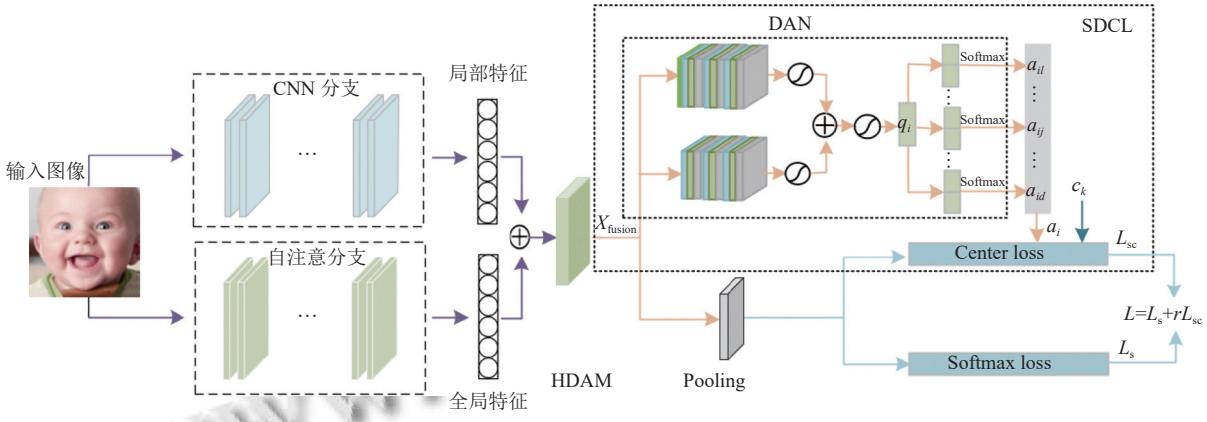


图 1 MIFNet 框架

将婴儿表情图像输入 MIFNet 之后, 由 CNN 分支和自注意力分支分别提取面部表情的局部细节特征与全局特征, 然后通过 HDAM 融合模块将两者在空间域与通道域进行融合, 最后利用自适应深度中心损失与 Softmax 损失联合估计相应婴儿表情类别的概率分布, 以确定婴儿图像的最终表情类别.

2.1 双分支特征提取结构

本文设计一个双分支特征提取结构, 其中 CNN 分支采用 ResNet18[18] 的骨干网络, 同时自注意力分支采用 Swin Transformer 的骨干网络, 图 2 展示了双分支特征提取结构每个阶段提取到的婴儿面部表情特征, 以及经过特征融合模块之后获得的融合特征. CNN 分支提取的特征更加突出局部细节特征, 缺失全局特征的信息. 自注意力分支提取的特征建立各像素之间的全局依赖性, 但缺少对于局部信息的关注, 不能准确地捕捉局部细节信息.

本文结合 CNN 结构与自注意力结构的特点, 利用 CNN 分支和自注意力分支分别提取婴儿表情局部

特征与全局特征用于网络后续的处理.

2.2 特征融合模块

根据自注意力机制的原理, Zhang 等人[19] 将多维度的特征分别进行降维处理, 融和降维特征应用于图像生成任务. 本文参考降维融合的思想提出一个空间全局注意力模块 (spatial global attention net, SGA-Net) 应用于婴儿面部表情识别任务之中. SGA-Net 通过融合降维生成的特征图获取输入特征的空间权重信息, 用以指导特征权重的分配, 其结构如图 3 所示.

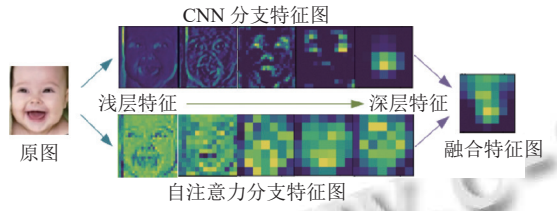


图 2 特征图可视化

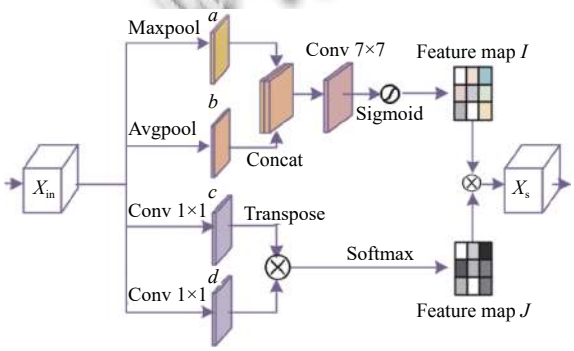


图 3 SGA-Net 模块

SGA-Net 模块将 CNN 分支提取的婴儿表情局部特征和自注意力分支提取的全局特征相加得到的特征 $X_{in} \in \mathbb{R}^{H\times W\times C}$ 通过支路 a 的全局最大池化操作获得特征每个通道向量中的最大元素值来表示特征信息, 同理 X_{in} 通过支路 b 的全局平均池化操作获得特征每个通道向量中的平均元素值来表示特征信息, 将二者拼接得到特征 X_{in} 在空间域的一种表示, 利用 7×7 卷积增强拼接特征, 最终通过 Sigmoid 函数进行激活得到融合特征 I, 详细计算过程为式 (2). 同时 X_{in} 通过支路 c 和支路 d 的 1×1 卷积进行通道变换, 实现对婴儿面部表情特征的降维处理转换成一维向量, 将支路 c 获取的降维特征转置之后再乘以支路 d 获取的降维特征, 用于融合表情特征的低维信息, 最终通过 Softmax 函数得到归一化特征 J, 详细计算过程为式 (3). 特征 J 转置之后乘以特征 I, 促进两者之间信息的交互, 并将获得的特征恢复至婴儿面部表情原始输入特征的维度, 使特征在空间域的信息得到增强.

$$I = S\left(f^{7\times 7}[AvgPool(X_{in}); MaxPool(X_{in})]\right) \tag{2}$$

其中, S 代表激活函数 Sigmoid, $f^{7\times 7}$ 代表 7×7 的卷积, $[\cdot]$ 代表矩阵的拼接, $AvgPool$ 代表全局平均池化操作, $MaxPool$ 代表全局最大池化操作.

$$J = \left(f^{1\times 1}\left(X_{in}^{C\times(W\times H)}\right) * f^{1\times 1}\left(X_{in}^{(W\times H)\times C}\right)\right) \tag{3}$$

其中, δ 代表 Softmax 函数, $f^{1\times 1}$ 代表 1×1 的卷积, * 代表矩阵相乘.

结合通道注意力机制的特点, 本文采用通道注意力模块 SK-Net (selective kernel network)[20] 用于特征通道信息的处理. SK-Net 从多尺度特征的视角出发, 利用两个不同卷积核的分支在不同尺度下学习特征, 获取特征重要尺度信息的权重, 增强网络对重要通道特征的关注, 其结构如图 4 所示.

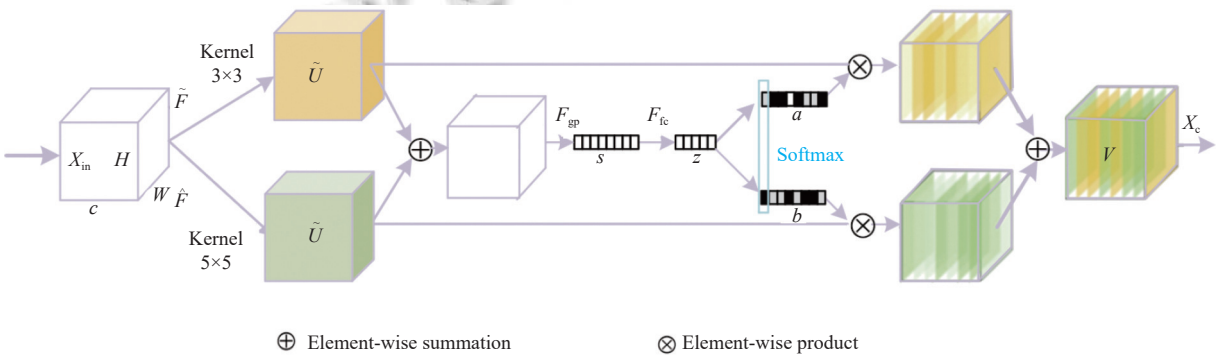


图 4 SK-Net 模块

SK-Net 模块将婴儿表情的局部特征和全局特征相加得到的特征 $X_{in} \in \mathbb{R}^{H\times W\times C}$ 同时输入 3×3 的空洞卷积 (感受野为 5×5) 与 3×3 的分组卷积得到特征 $\tilde{U} \in \mathbb{R}^{H'_d\times W'_d\times C'_f}$ 与特征 $\hat{U} \in \mathbb{R}^{H'_d\times W'_d\times C'_f}$, 为了达到自适应调整

感受野范围的效果, 通过门控单元控制多分支的信息流, 将多分支信息整合得到$U \in \mathbb{R}^{H'_d \times W_d \times C'_f}$. 整合后的特征$U$经过全局平均池化得到的特征图$s \in \mathbb{R}^{1\times1\times C}$, 再通过$\mathcal{F}_{\rm fc}(s)$函数生成向量$z \in \mathbb{R}^{d\times1}$, 并由式 (4) 和式 (5) 计算得到$a_c \in \mathbb{R}^{C\times1}$和$b_c \in \mathbb{R}^{C\times1}$

$$a_c = \frac{e^{A_c \times z}}{e^{A_c \times z} + e^{B_c \times z}} \tag{4}$$

$$b_c = \frac{e^{B_c \times z}}{e^{A_c \times z} + e^{B_c \times z}} \tag{5}$$

其中, a_c和b_c中的矩阵$A_c \in \mathbb{R}^{1\times d}$和$B_c \in \mathbb{R}^{1\times d}$在训练之前需要进行初始化.

a_c和b_c的函数值和为 1, 在对分支的特征图权重进行设置时, 控制两条分支的比重. 将a_c与$\tilde{U}_c$, b_c与$\hat{U}_c$相乘得$V_c \in \mathbb{R}^{H\times W}$, 整合获得特征图$V \in \mathbb{R}^{H\times W\times C}$.

$$V_c = a_c \cdot \tilde{U}_c + b_c \cdot \hat{U}_c \tag{6}$$

$$V = [V_1, V_2, \cdots, V_c] \tag{7}$$

其中, a_c和b_c是$\tilde{U}$和$\hat{U}$分别对应的注意力向量, A_c与B_c分别代表A与B的第 c 行, a_c与b_c分别代表 a 与 b 的第 c 个元素, $[\cdot]$ 代表矩阵的拼接.

在空间注意力模块中, 特征的各通道被等同考虑, 忽视通道之间的信息交流; 在通道注意力模块中, 则是直接对特征的每个通道进行整体分析, 且不考虑空间信息之间的交流. 因此本文将空间域注意力模块 SGA-Net 和通道域注意力模块 SK-Net 并联得到一个混合域注意力模块 (hybrid domain attention module, HDAM), 其结构如图 5 所示. 在通道域与空间域分别处理特征, 保留特征通道与空间的关键信息, 增强特征的表达能力. 本文通过实验证明, SGA-Net 与 SK-Net 并行的连接方式比 SK-Net 与 SGA-Net 串联的方式效果更佳.

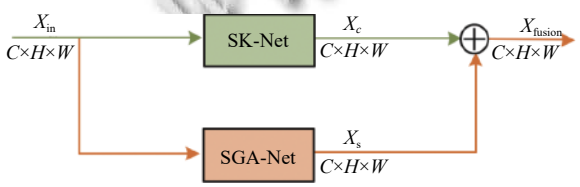


图 5 HDAM 框架

2.3 自适应深度中心损失

本文针对婴儿面部表情类间相似性高的特点, 设计了一个自适应深度中心损失, 利用注意力网络降低不同类别表情之间共同特征的权重, 提高不同类别表情差异特征的权重, 联合 Softmax 损失对网络进行优化, 实现婴儿面部表情的精准识别. 具体计算如式 (8) 所示.

$$L = L_{\rm S} + \gamma L_{\rm SC} \tag{8}$$

其中, 平衡系数γ设为 0.6, $L_{\rm S}$为 Softmax 损失, $L_{\rm SC}$为自适应深度中心损失.

MIFNet 网络采用的损失函数如图 6 所示, 双分支结构提取的特征经 HDAM 模块后获得融合特征, 通过一个特征池化层提取最终的 d 维深度特征向量, 将其用于 Softmax 损失和自适应深度中心损失进行表情类别判断. 同时作为上下文输入到注意力网络用于估计注意力权重, 并将得到的权重用于指导中心损失, 实现特征元素子集中婴儿表情特征的类内紧凑与类间分离, 联合 Softmax 损失进行优化.

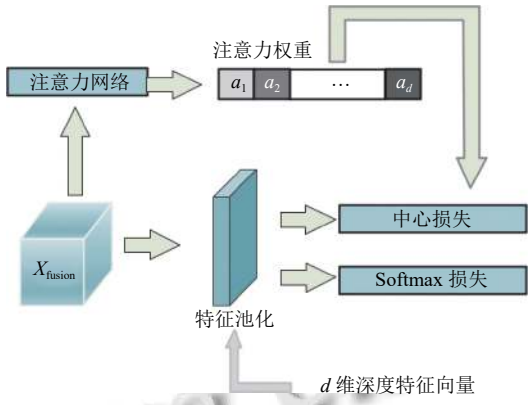


图 6 MIFNet 的损失框架

在一个特征向量中并非所有的元素都与分类有关, 本文目标是在婴儿面部表情的深度特征向量中减少不相关的元素, 从而获得向量中一个子集的元素来辅助分类. 为了在分类过程中过滤掉不相关的特征和婴儿不同表情类别的共同特征, 本文在中心损失计算方法的基础上对每个维度的欧氏距离进行加权计算, 提出自适应深度中心损失 (self-adaptive deep centre loss, SDCL), 具体计算如式 (9) 所示.

$$L_{\rm SC} = \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{d} a_{ij} \odot \| x_{ij} - c_{y_{ij}} \|_2^2 \tag{9}$$

其中, $\odot$代表元素相乘, a_{ij} 代表在嵌入空间中由深度注意网络计算得出的沿维度$j \in \{1,\cdots,d\}$的第 i 个深度特征的权重.

本文设计了一种自适应且灵活的方法来估计输入

特征的权重, 以适应婴儿面部表情识别任务, 为此本文设计了一个深度注意网络 (deep attention network, DAN) 连接到 HDAM 模块之后, 根据输入的婴儿面部表情特征动态地估计自适应深度中心损失的注意力权重 a_{ij} 指导中心损失. 深度注意网络的结构如图 7 所示, 主要包含两个组成部分: 上下文编码模块 (context encoder moduel, CEU), 该模块将 HDAM 模块获得的融合特征图作为输入生成一个潜在表征向量; 多头二元分类模块 (multi-head binary classification moduel, MBCM), 该模块利用潜在表征向量估计注意力权重. 本文的深度注意网络对特征的处理是在卷积特征层面, 因此保留婴儿表情特征的空间信息.

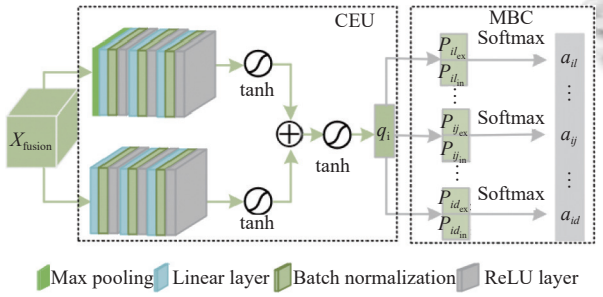


图 7 深度注意网络

深度注意网络的双分支上下文编码模块上分支结构通过堆叠全局最大池化层和 3 个可训练的全连接线性层构成, 计算如式 (10) 所示; 下分支结构通过堆堆叠 3 个可训练的全连接线性层构成, 计算如式 (11) 所示.

$$q_{i1}=\tanh(BN(W_3^{\mathrm{T}}ReLU(BN(W_2^{\mathrm{T}}ReLU(BN(\cdots W_1^{\mathrm{T}}\alpha(X_{\mathrm{fusion}})+b_1))+b_2))+b_3)) \tag{10}$$

$$q_{i2}=\tanh(BN(W'^{\mathrm{T}}_3ReLU(BN(W'^{\mathrm{T}}_2ReLU(BN(\cdots W'^{\mathrm{T}}_1\beta(X_{\mathrm{fusion}})+b'_1))+b'_2))+b'_3)) \tag{11}$$

其中, X_{fusion} 代表输入的婴儿表情图像在通过 HDAM 模块后获得的融合特征图, 即第 i 个样本的上下文特征. 算子α: 将特征图 X_{fusion} 进行全局最大池化和展平处理. 算子β: 将特征图 X_{fusion} 进行展平处理. W_l和b_l分别代表上下文编码模块上分支的第 l 个线性层的权重和偏置, 其中, $l=1,2,3$. W'_l和b'_l代表上下文编码模块下分支的第 l 个线性层的权重和偏置, 其中, $l=1,2,3$.

深度注意网络的上下文编码模块各网络层之间插入批量归一化层 $BN(\cdot)$ 和整流的线性单元 $ReLU(\cdot)$, 其目的是捕捉层间非线性关系, $\tanh(\cdot)$ 作为激活函数保留正负激活值, 使网络中的梯度流更加平滑. 通过式 (8) 与式 (9) 分别计算得到上下文编码模块上分支与下分支第 i 个样本的的潜在特征向量$q_{i1}\in\mathbb{R}^{d'<<d}$和$q_{i2}\in\mathbb{R}^{d'<<d}$, 并通过式 (12) 计算得到低纬度的潜在特征向量 $q_i\in\mathbb{R}^{d'<<d}$, 用以消除不相关信息, 保留重要特征信息. 上下文编码模块可以根据层级参数进行调整, 实现匹配到特定任务之中.

$$q_i=\tanh(q_{i1}+q_{i2}) \tag{12}$$

深度注意网络的多头二元分类 (包含/排除) 模块位于上下文编码模块之后, 潜在的d'维特征向量q_i在d个线性单元之间共享, 即每个单元均有两个输出来计算深度特征 X_{fusion} 沿维度 j 的两个原始分数, 计算过程如式 (13) 所示.

$$\begin{cases} p_{ij_{\mathrm{in}}}=A_{j_{\mathrm{in}}}^{\mathrm{T}}\mathrm{e}_i+b_{j_{\mathrm{in}}} \\ p_{ij_{\mathrm{ex}}}=A_{j_{\mathrm{ex}}}^{\mathrm{T}}\mathrm{e}_i+b_{j_{\mathrm{ex}}} \end{cases} \tag{13}$$

其中, $A_j\in\mathbb{R}^{d'\times 2}$和$b_j\in\mathbb{R}^2$代表多头二元分类模块每个分类头的可学习权重和偏差, 下标 in 代表包含, 下标 ex 代表排除, $p_{ij_{\mathrm{in}}}$和$p_{ij_{\mathrm{ex}}}$分别代表 X_{fusion} 中第 j 维度的包含分数和排除分数.

将式 (11) 计算得到的两个原始分数使用 Softmax 函数归一化每个头的输出, 得出每个维度相应的注意力权重, 即自适应深度中心损失的注意力权重 a_{ij}, 在原始分数上采用 Softmax 函数进行微调, 将估计的注意权重值限制在 (0, 1) 范围内, 计算过程如式 (14) 所示.

$$a_{ij}=\frac{\exp(p_{ij_{\mathrm{in}}})}{\exp(p_{ij_{\mathrm{in}}})+\exp(p_{ij_{\mathrm{ex}}})} \tag{14}$$

3 数据集介绍及模型训练

3.1 婴儿面部表情数据集的构建

国内外现存公开的婴儿面部表情数据集包括 TIF 数据集[21] 和 CIF 数据集[22]. 其中 TIF 数据集将婴儿表情分为: 快乐, 悲伤, 厌恶和平静 4 类, 共包含 119 张图像. CIF 数据集将婴儿表情分为: 积极, 消极和平静 3 类, 共包含 154 张图像. 以上两个数据集的图像均不是在自然环境中拍摄, 图像经过特殊处理且样本数量太少, 因此不适用于深度学习方法.

本文从网上收集 10 240 张的婴儿 (1 岁以内) 图像, 建立婴儿面部表情数据集 IFER (infant facial expression recognition). 采用 RetinaFace[23] 检测婴儿面部的

5 个关键点, 再通过仿射变换对齐面部, 最后将图像裁剪至 256×256 像素, 样本示例如图 8 所示.

图 8 IFER 数据集各类表情图像示例

根据公开数据集 TIF 数据集与 CIF 数据集对于婴儿表情的分类, 本文将 IFER 表情分为 4 类: 快乐、平静、哭泣、睡眠. 在专业的心理医师与儿科医护人员指导下, 从描绘的表情、表情的清晰度和表达的强度 3 方面对图像中婴儿表情进行评分, 判定其所属类别. 将数据集 IFER 的每类婴儿表情图像按照 6:2:2 划分为训练集、验证集和测试集, 划分细节如表 1 所示.

表 1 数据集划分情况

表情类别	样本数量			
	训练集	验证集	测试集	合计
快乐	1 946	648	648	3 242
平静	2 186	728	728	3 642
哭泣	1 195	398	398	1 991
睡眠	819	273	273	1 365

3.2 RAF-DB 数据集

面部表情公开数据集 RAF-DB (real-world affective faces database)[24] 是一个大规模的无约束表情数据集, 包含 29 672 张表情图像. 所有图像均是在互联网上收集, 研究对象的年龄、性别、种族存在很大差异. 该数据集包含两个部分: 单标签子集和双标签子集, 其中, 单标签子集分为 7 类基本情绪: 愤怒、厌恶、恐惧、快乐、悲伤、惊讶和中性; 双标签子集包含 2 种基本情绪的组合, 共 12 类复合情绪. 单标签子集共包含 15 339 张表情图像, 其中 12 271 张图像为训练数据, 3 068 张图像为测试数据, 图像分辨率为 100×100.

3.3 模型训练

本文在 Ubuntu 16.04 操作系统下, 通过 PyTorch 1.2.0 框架实现 MIFNet 的训练与测试, 硬件环境为 NVIDIA-GTX 1080Ti GPU, 显存为 11 GB. MIFNet 网络参数通过标准的随机梯度下降法进行优化, 动量设置为 0.9, 权重衰减设置为 0.000 5, 初始学习率设置为 0.01, batch size 设置为 64.

4 实验结果与分析

4.1 实验评价指标

对于单标签任务而言, 网络模型预测结果和标签值一致则认为分类正确, 不一致则认为分类错误. 因此对于本文的婴儿面部表情识别任务而言, 最直观的评估指标是准确率. 准确率表示预测正确的样本占总样本的百分比, 其计算过程如式 (15) 所示.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{15}$$

其中, 真阳性 TP (true positive) 表示网络模型将正类样本判断为正类的数量, 假阳性 FP (false positive) 表示网络模型将负类样本判断为正类的数量, 真阴性 TN (true negative) 表示网络模型将负类样本判断为负类的数量, 假阴性 FN (false negative) 表示网络模型将正类样本判断为负类的数量.

除准确率之外, 本文同时使用 AUC (area under curve)、召回率 (*recall*)、$F1$ 分数 3 个指标定量分析模型的性能.

AUC (受试者操作特征) 指的是 ROC (receiver operating characteristic curve) 曲线下的面积, 其中 ROC 曲线的横轴为假阳性率 (false positive rate, FPR), 纵轴为真阳性率 (true positive rate, TPR). AUC 值越大, 当前的婴儿表情分类算法越有可能将正样本排在负样本前面, 即具有更好的分类能力, 其计算过程如下.

$$FPR = \frac{FP}{FP+TN} \tag{16}$$

$$TPR = \frac{TP}{TP+FN} \tag{17}$$

召回率用于计算所有婴儿表情的真实标签为阳性的样本中真阳性的比例. 较高的召回率意味着模型识别阳性样本的能力更强, 其计算过程如下.

$$recall = \frac{TP}{TP+FN} \tag{18}$$

$F1$ 分数是精确率和召回率两个指标的组合, $F1$ 分数这个综合评价指标可以用来评价婴儿表情网络模型的整体性能, $F1$ 值越高表示分类能力越强.

$$F1=\frac{2TP}{2TP+FP+FN} \tag{19}$$

4.2 MIFNet 网络结构消融实验

为了验证本文设计的双分支特征提取结构、特征融合模块和自适应深度中心损失的有效性, 本文基于婴儿面部表情数据集 IFER 对 MIFNet 进行消融实验, 并作出定量分析. 通过控制变量法进行如下实验: (1) MIFNet 只使用 ResNet18 的骨干网络用于特征提取; (2) MIFNet 只使用 Swin Transformer 的骨干网络用于特征提取; (3) MIFNet 使用双分支特征提取结构用于特征提取; (4) MIFNet 使用双分支特征提取结构与特征融合模块用于特征提取但采用交叉熵损失替换本文损失用于婴儿表情分类. 实验结果如表 2 所示.

表 2 MIFNet 在 IFER 上的消融实验结果 (%)

序号	骨干网络	HDAM	SDCL	快乐	平静	哭泣	睡眠	平均准确率	标准准确率
①	Swin	×	√	92.93	91.61	96.53	98.79	94.97	93.95
②	Res	×	√	93.38	91.81	96.42	98.41	95.01	94.09
③	Res+Swin	×	√	93.95	92.34	96.79	99.12	95.55	94.61
④	Res+Swin	√	×	93.68	91.93	96.59	98.94	95.29	94.35
⑤	Ours	√	√	**94.44**	**93.54**	**97.74**	**99.63**	**96.34**	**95.46**

由于数据集 IFER 的表情类别不平衡, 因此本文在表 2 中列出消融实验获得的标准准确率、平均准确率和各类别的准确率, 其中平均准确率为婴儿表情各类别准确率平均值. 通过表 2 可以直观看出 MIFNet 各组成部分对网络总体性能的影响. 对①、②、和③进行分析, 自注意力分支结构关注全局信息并考虑每个像素点之间的联系, 可以提取到婴儿面部表情的全局特征, 同时 CNN 分支可以提取到婴儿面部表情的局部特征, 因此采用双分支特征提取结构获得整体性特征的方法优于每个分支单独提取特征的方法. 对③和⑤进行分析, HDAM 模块加入双分支特征提取结构之后, 将提取的婴儿表情特征沿着通道域与空间域进行特征融合, 使婴儿面部表情识别的标准准确率提升 0.85%, 效果显著. 对④和⑤进行分析, 在 MIFNet 的特征提取结构与融合模块保持不变的情况下, 将本文设计的损失与交叉熵损失相比较, 本文损失使婴儿面部表情识别的标准准确率提升 1.11%.

4.3 损失函数有效性实验

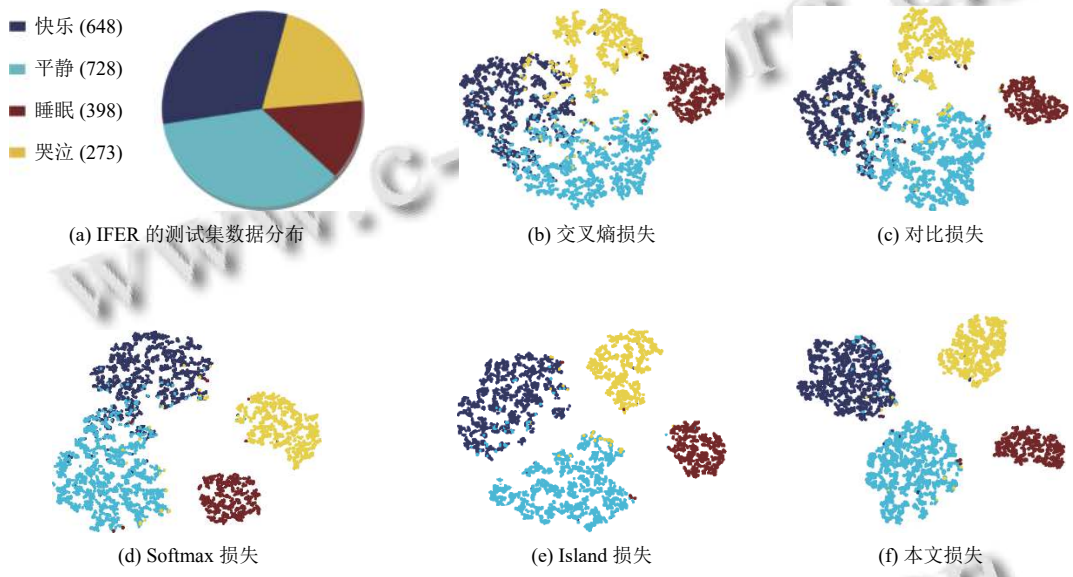
除网络结构设计之外, 损失函数同样影响婴儿面部表情识别的准确率, 在本文任务的网络训练过程中, 选择恰当的损失函数将预测值与婴儿表情真实标签值之间的误差进行反向传播来指导 MIFNet 模型的迭代优化. 本文针对婴儿面部表情的特点, 设计自适应深度中心损失, 并采用自适应深度中心损失与 Softmax 损失联合优化的方法, 在嵌入空间中促进婴儿同一表情类别的特征紧凑和不同类别间的特征分离. 为验证本文损失的有效性, 在保持 MIFNet 其余组成部分不变的前提下, 使用其他常见的分类损失替换 MIFNet 的损失在数据集 IFER 上进行一系列的对比实验, 实验结果如表 3 所示. 针对本文婴儿面部表情识别任务, MIFNet 采用本文损失获得的召回率、$F1$ 分数和准确率 3 个指标均为最高. 针对婴儿面部表情类间相似性高的问题, 本文损失在嵌入空间中促进了深度特征的类内紧凑与类间分离, 效果表现优异. 但本文损失是由两个损失函数组成的联合损失, 在网络训练过程中收敛速度较慢.

表 3 不同损失函数在 IFER 上的对比结果 (%)

损失函数	准确率	召回率	$F1$分数	AUC
交叉熵损失	94.31	94.94	95.17	98.85
对比损失	94.45	95.51	95.40	99.03
Softmax损失	94.63	95.51	95.44	99.01
Island损失	94.89	95.69	95.71	**99.08**
本文损失	**95.46**	**95.88**	**95.89**	99.07

为了更加形象化地看出各损失函数在本文分类任务上的性能, 如图 9 所示本文利用 t-SNE 算法可视化了各损失函数的聚类结果, 其中 IFER 的测试集中各表情类别的数据分布如图 9(a) 所示. 交叉熵损失常用于分类任务, 针对本文的分类任务, 交叉熵损失注重婴儿不同类别表情之间相差较大的特征, 缺乏对类间相似度的关注, 其聚类效果如图 9(b) 所示, 在本文所对比的损失中效果最差. 对比损失聚集具有相同标签的表情深度特征, 分离不同标签的表情深度特征, 将对比损失迁移到本文的婴儿面部表情识别任务中, 由于婴儿表

情相对成人表情具有更高的类间相似性, 导致其对婴儿表情深度特征的分离效果下降, 其聚类效果如图 9(c) 所示. Softmax 损失将学习到的特征在深度空间形成表情类别簇, 通过对分类错误的婴儿面部表情进行惩罚来保持不同表情类别之间分离, 但 Softmax 损失形成的每个表情类别簇的特征是分散的, 并未实现婴儿同一表情类别特征紧凑的目的, 其聚类效果如图 9(d) 所示. Island 损失应用于婴儿面部表情识别任务, 相对 Softmax 损失压缩了每个表情簇的簇内间距并将类中心作为孤立的岛屿分开, 促使婴儿同一表情类别间的特征紧凑, 分类效果有了相对的提升, 其聚类效果如图 9(e) 所示. 本文提出的自适应深度中心损失关注到婴儿表情类间相似性高的问题, 通过对表情特征向量元素赋予不同的权重削弱无关元素的影响, 本文损失采用深度注意损失与 Softmax 损失联合优化的方法, 其聚类效果如图 9(f) 所示. 综合分析, 针对婴儿面部表情识别的任务, 本文损失在嵌入空间中实现了最优的类内紧凑与类间分离效果.



(a) IFER 的测试集数据分布 (b) 交叉熵损失 (c) 对比损失

(d) Softmax 损失 (e) Island 损失 (f) 本文损失

图 9 各损失聚类效果图

4.4 基于数据集 IFER 的实验结果与分析

为了验证 MIFNet 的有效性, 本文将 MIFNet 与近几年面部表情识别网络在数据集 IFER 上进行了一系列的对比实验, 不同方法在数据集 IFER 上获得的标准准确率和平均准确率如表 4 所示, 不同方法在数据集 IFER 上的混淆矩阵如图 10 所示.

表 4 MIFNet 与其他网络在 IFER 上的对比结果

方法	年份	平均准确率 (%)	标准准确率 (%)
IRA2LT[25]	2018	91.93	91.35
RAN[11]	2019	93.62	92.79
SCN[13]	2020	94.64	93.63
DACL[26]	2021	94.96	93.94
RUL[27]	2021	95.03	94.02
DAN[28]	2021	95.13	94.21
Ad-Corre[29]	2022	94.31	93.40
MA-Net*[30]	2022	95.81	94.87
Ours	2022	**96.34**	**95.46**

实验结果表明, MIFNet 在数据集 IFER 上的表现优于近几年的面部表情识别网络, 标准准确率和平均准确率分别是 95.46%、96.34%, 均高于其他网络. 综合分析原因如下, 近几年的面部表情识别网络研究对象包含各年龄段的人群, 个体面部在不同年龄阶段有不一样的表现, 因此网络更加关注于不同人群同一表情类别的共同特征. 婴儿面部线条流畅、五官紧凑且锐感偏弱, 而成人的面部轮廓清晰、五官立体且敏锐, 导致婴儿表情类间相似性高, 即使是不同表情类别之间也存在很多共同特征, 因此近几年的面部表情识别网络对于婴儿表情的类间分离效果不理想, 容易造成表情误检. 本文设计的网络不仅可以提取婴儿面部表情整体性的特征, 并且解决了婴儿表情类间相似性高的问题, 因而准确率达到最高.

4.5 基于公开数据集 RAF-DB 的实验结果与分析

RAF-DB 是通过互联网收集图像创建的数据集,

与在实验室收集的数据集相比, RAF-DB 存在更多嘈杂的标签和噪声, 所以其对于表情识别任务更具有挑战性. 为了验证 MIFNet 的泛化性, 本文选择 RAF-DB 的单标签子集对 MIFNet 进行实验分析, MIFNet 网络与其他网络在 RAF-DB 上对比结果如表 5 所示.

实验结果表明, 基于 RAF-DB 数据集将 MIFNet 与近几年表情识别网络相比, MIFNet 的准确率处于第 2 位. 相对于婴儿数据集, RAF-DB 数据集中的图像具有更多嘈杂的噪声, 而 MIFNet 对于噪声的关注低于 MA-Net*, 导致其准确率低于 MA-Net*. 但针对本文婴儿表情识别任务的特点, MIFNet 更加关注婴儿表情类间相似性高的特点, 识别效果明显优于 MA-Net*.

图 10 数据集 IFER 基于不同方法的混淆矩阵

以往面部表情识别的网络主要是以 CNN 为骨干网络, 可以获取到面部表情的局部特征, 仅靠局部特征分辨表情类别具有一定的局限性. 本文将 Swin Transformer 的骨干网络作为自注意力分支来建立面部表情特征之间的全局联系, 减少全局信息的丢失, 同时利用融合模块把 CNN 分支关注的局部特征融入全局特征, 在全局特征中更加突出重点区域. 同时本文提出的自适应深度中心损失实现了将 7 类表情深度特征在嵌入空间中进行分离, 并提高每类表情特征间的紧凑性. 基于上述原因, MIFNet 在 RAF-DB 数据集上表现良好.

表 5 MIFNet 与其他网络在 RAF-DB 上对比结果

方法	年份	平均准确率 (%)	标准准确率 (%)
IRA2LT[25]	2018	—	86.77
RAN[11]	2019	—	86.90
SCN[13]	2020	—	87.03
DACL[26]	2021	80.44	87.78
EfficientFace[31]	2021	—	88.36
RUL[27]	2021	—	88.98
DAN[28]	2021	85.32	89.70
MA-Net*[30]	2022	—	**89.99**
Ours	2022	86.29	89.87

5 结论与展望

本文基于面部表情识别的研究现状, 针对婴儿表情的特点提出一个分类网络 MIFNet 用于识别婴儿面部表情. 近几年的面部表情识别网络大部分使用 CNN 作为特征提取网络, 获得的特征具有局限性, MIFNet 利用混合域融合模块将 CNN 分支提取的局部特征与自注意力分支提取的全局特征进行融合实现对婴儿面部表情多尺度特征的提取, 从而获得婴儿面部表情的整体性特征. 针对婴儿面部表情类间相似性高的问题, 本文设计了一个自适应深度中心损失, 联合 Softmax 损失对网络进行优化, 提高了网络的分类能力. 在 IFER 数据集中 MIFNet 实现对婴儿面部表情精准识别, 效果优于目前现有的深度学习方法. 如果能将网络获取的特征转换成数据指标解析婴儿的状态, 可进一步判断婴儿身心健康, 因此未来的研究将关注如何将神经网络提取的婴儿表情特征转换为数据指标分析.

参考文献

1 Fang CY, Ma CW, Chiang ML, *et al*. An infant emotion recognition system using visual and audio information. Proceedings of the 4th International Conference on Industrial Engineering and Applications (ICIEA). Nagoya: IEEE, 2017. 284–291.

2 Zhang LR, Xu C, Li S. Facial expression recognition of infants based on multi-stream CNN fusion network. Proceedings of the 5th International Conference on Signal and Image Processing. Nanjing: IEEE, 2020. 37–41.

3 Zamzami G, Ruiz G, Goldgof D, *et al*. Pain assessment in infants: Towards spotting pain expression based on infants' facial strain. Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Ljubljana: IEEE, 2015. 1–5.

4 Messinger DS, Mahoor MH, Chow SM, *et al*. Automated measurement of facial expression in infant-mother interaction: A pilot study. Infancy, 2009, 14(3): 285–305. [doi: 10.1080/15250000902839963]

5 Matsugu M, Mori K, Mitari Y, *et al*. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks, 2003, 16(5–6): 555–559.

6 李小薪, 梁荣华. 有遮挡人脸识别综述: 从子空间回归到深度学习. 计算机学报, 2018, 41(1): 177–207. [doi: 10.11897/SP.J.1016.2018.00177]

7 林景栋, 吴欣怡, 柴毅, 等. 卷积神经网络结构优化综述. 自动化学报, 2020, 46(1): 24–37.

8 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

9 Liu Z, Lin YT, Cao Y, *et al*. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.

10 徐玮, 郑豪, 杨种学. 基于双注意力模型和迁移学习的 Apex 帧微表情识别. 智能系统学报, 2021, 16(6): 1015–1020. [doi: 10.11992/tis.202010031]

11 Wang K, Peng XJ, Yang JF, *et al*. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 2020, 29: 4057–4069. [doi: 10.1109/TIP.2019.2956143]

12 Gera D, Balasubramanian S. Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. Pattern Recognition Letters, 2021, 145: 58–66. [doi: 10.1016/j.patrec.2021.01.029]

13 Wang K, Peng XJ, Yang JF, *et al*. Suppressing uncertainties for large-scale facial expression recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6896–6905.

14 Yu NG, Bai DG. Facial expression recognition by jointly partial image and deep metric learning. IEEE Access, 2020, 8: 4700–4707. [doi: 10.1109/ACCESS.2019.2963201]

15 Wen YD, Zhang KP, Li ZF, *et al*. A discriminative feature learning approach for deep face recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 499–515.

16 Cai J, Meng ZB, Khan AS, *et al*. Island loss for learning discriminative features in facial expression recognition.

Proceedings of 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Xi'an: IEEE, 2018. 302–309.

17 Li YJ, Lu Y, Li JX, *et al*. Separate loss for basic and compound facial expression recognition in the wild. Proceedings of the 11th Asian Conference on Machine Learning. Nagoya: PMLR, 2019. 897–911.

18 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

19 Zhang H, Goodfellow IJ, Metaxas DN, *et al*. Self-attention generative adversarial networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 7354–7363.

20 Li X, Wang WH, Hu XL, *et al*. Selective kernel networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 510–519.

21 Maack JK, Bohne A, Nordahl D, *et al*. The Tromso infant faces database (TIF): Development, validation and application to assess parenting experience on clarity and intensity ratings. Frontiers in Psychology, 2017, 8: 409. [doi: 10.3389/fphys.2017.00409]

22 Webb R, Ayers S, Endress A. The city infant faces database: A validated set of infant facial expressions. Behavior Research Methods, 2018, 50(1): 151–159. [doi: 10.3758/s13428-017-0859-9]

23 Deng JK, Guo J, Ververas E, *et al*. RetinaFace: Single-shot multi-level face localisation in the wild. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5202–5211.

24 Li S, Deng WH, Du JP. 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. Honolulu: IEEE, 2017. 2584–2593.

25 Zeng JB, Shan SG, Chen XL. Facial expression recognition with inconsistently annotated datasets. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 227–243.

26 Farzaneh AH, Qi XJ. Facial expression recognition in the wild via deep attentive center loss. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 2401–2410.

27 Zhang YH, Wang CR, Deng WH. Relative uncertainty learning for facial expression recognition. Proceedings of the 35th Conference on Neural Information Processing Systems. 2021. 17616–17627.

28 Wen ZY, Lin WZ, Wang T, *et al*. Distract your attention: Multi-head cross attention network for facial expression recognition. arXiv:2109.07270, 2021.

29 Fard AP, Mahoor MH. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access, 2022, 10: 26756–26768. [doi: 10.1109/ACCESS.2022.3156598]

30 Heidari N, Iosifidis A. Learning diversified feature representations for facial expression recognition in the wild. arXiv:2210.09381, 2022.

31 Zhao ZQ, Liu QS, Zhou F. Robust lightweight facial expression recognition network with label distribution training. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI Press, 2021. 3510–3519.

(校对责编: 牛欣悦)