通道-空间多尺度增强与双池化注意的表情识别网络*

刘 娟[1†], 张民扬[1], 胡 敏[2], 黄 忠[1,2], 江巨浪[1]
(1. 安庆师范大学 电子工程与智能制造学院, 安徽 安庆 246133; 2. 合肥工业大学 计算机与信息学院 情感计算与先进智能机器安徽省重点实验室, 合肥 230009)

摘 要: 针对自然场景下表情特征提取仅关注通道-空间单一尺度信息以及平均池化易丢失局部显著性语义的问题,提出一种通道-空间多尺度增强与双池化注意的表情识别网络。首先,为捕获通道-空间整体多尺度增强语义,设计通道对称级联多尺度模块和空间多尺度特征提取模块,并以此构建基于通道-空间多尺度结构的整体特征增强子网。然后,为表征通道-空间区域双池化显著语义,将高效局部注意力机制改进为高效通道-空间注意力机制,并嵌入到区域特征注意子网。最后,为获取整体多尺度增强语义与区域双池化显著语义之间的潜在相关性,采用交叉注意力机制进行整体特征与区域特征之间的特征交互,并设计特征融合子网完成两类特征的模型级融合。实验结果表明,在人脸表情数据集 RAF-DB 和 FERPlus 上,其表情识别率分别达到 89.97% 和 90.26%,比基线网络分别提升了 13.54 和 10.95 个百分点。与其他网络相比,提出的网络在自然场景下具有较好的表情识别性能。

关键词: 人脸表情识别; 多尺度增强; 双池化注意; 通道-空间多尺度结构; 高效通道-空间注意力机制
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 1001-3695(2025)10-037-3182-10
doi: 10.19734/j.issn.1001-3695.2024.12.0524

Expression recognition network of channel-spatial multi-scale enhancement and dual-pooling attention

Liu Juan[1†], Zhang Minyang[1], Hu Min[2], Huang Zhong[1,2], Jiang Julang[1]
(1. *School of Electronic Engineering & Intelligent Manufacturing, Anqing Normal University, Anqing Anhui* 246133, *China*; 2. *Anhui Province Key Laboratory of Affective Computing & Advanced Intelligent Machine, School of Computer Science & Information Engineering, Hefei University of Technology, Hefei* 230009, *China*)

Abstract: Aiming at the problems that expression feature extraction in natural scenes only focuses on channel-spatial single-scale information and average pooling is easy to lose local saliency semantics, this paper proposed an expression recognition network of channel-spatial multi-scale enhancement and dual-pooling attention. Firstly, to capture the whole channel-spatial multi-scale enhancement semantics, this paper designed a channel symmetric cascade multi-scale module and a spatial multi-scale feature extraction module, and constructed a whole feature enhancement subnetwork based on the channel-spatial multi-scale structure. Then, to represent the channel-spatial region dual-pooling salient semantics, this paper improved the efficient local attention mechanism into an efficient channel-spatial attention mechanism, and embedded it into the region feature attention subnetwork. Finally, to obtain the potential correlation between the whole multi-scale enhanced semantics and the regional dual-pooling salient semantics, this paper used the cross-attention mechanism to perform the feature interaction between the whole features and the regional features, and designed the feature fusion subnetwork to complete the model-level fusion of the two types of features. The experimental results show that the expression recognition rates on the facial expression datasets RAF-DB and FERPlus reach 89.97% and 90.26% respectively, which are 13.54 and 10.95 percentage points higher than the baseline network. Compared with other networks, the proposed network has better expression recognition performance in natural scenes.

Key words: facial expression recognition; multi-scale enhancement; dual-pooling attention; channel-spatial multi-scale structure; efficient channel-spatial attention mechanism

0 引言

面部表情是传达人类情感最有效方法和手段之一,在远程教育、人机交互、医疗诊断和辅助驾驶等自然场景中受到广泛关注[1,2]。然而,面部表情易受姿态变化、遮挡、光照等因素的影响[3,4],如何提升自然场景下表情识别的鲁棒性成为当前情感计算和“情智兼备”领域的研究热点[5,6]。

当前,自然场景下的表情识别方法大致可分为:基于面部整体语义、基于面部区域语义以及基于两者语义融合的表情识别方法[7,8]。基于面部整体语义的方法主要采用传统特征提取方法或注意力机制来获取面部整体特征的表示。如 Ma 等人[9]基于卷积神经网络和局部二值模式分别提取面部表情的

收稿日期: 2024-12-06; **修回日期**: 2025-02-16 **基金项目**: 国家自然科学基金资助项目(62176084);安徽省教育厅自然科学重点研究项目(2022AH051038,2023AH050500,2023AH050474)

作者简介: 刘娟(1984—),女(通信作者),安徽安庆人,副教授,硕导,硕士,主要研究方向为情感计算、计算机视觉(juanliu3039@163.com);张民扬(1998—)男,安徽合肥人,硕士,主要研究方向为情感计算、计算机视觉;胡敏(1967—),女,安徽淮北人,教授,博导,博士,主要研究方向为情感计算、图像处理;黄忠(1981—),男,安徽安庆人,教授,硕导,博士,主要研究方向为情感计算、自然人机交互;江巨浪(1967—),男,安徽安庆人,教授,硕导,博士,主要研究方向为图像处理、计算机视觉.

深度特征和纹理特征,并级联两类特征以获取面部全局语义。Li 等人[10]堆叠局部二值标准层以捕获不同尺度的纹理特征,并输入至注意力网络以获取面部全局特征。然而,基于面部整体语义的表情识别方法未能捕获面部局部区域的细节信息,且自然场景下头部姿态、面部遮挡等因素易降低全局特征语义的区分度[11]。与前者不同,基于面部区域语义的表情识别方法更加关注于面部局部信息。Wang 等人[12]采用自注意力机制计算重叠面部区域的权重,自适应地捕捉关键人脸区域信息。Huang 等人[13]将面部图像划分为不重叠的区域子块,并借助网格注意力机制提取区域特征。尽管这些方法对面部细节信息进行了深入挖掘,但对整体面部特征的完整性及不同表情区域之间协作关系的考虑较为不足[14]。为充分发挥整体语义和区域语义的优势,基于整体-区域语义融合的表情识别方法成为当前人脸表情识别研究的主流方法。Liu 等人[15]基于 ViT(vision Transformer)获取单一尺度的整体特征,并采用单池化操作自适应计算区域块的注意力权重以获取区域特征。Zhao 等人[16]通过对称 Res2Net[17]结构提取整体通道多尺度特征,并引入 CBAM(convolutional block attention module)[18]获取区域语义。Liu 等人[19]采用自适应多层感知注意网络提取多尺度全局通道特征和局部显著特征。Gera 等人[20]采用空间通道注意网获取全局上下文特征,并基于高效通道注意力捕获局部特征。Xiao 等人[21]根据全局和局部面部信息的重要性自动分配相应的权重,并将全局特征和局部特征进行特征级融合。He 等人[22]通过高效集成模块和局部提取模块分别捕获整体特征和区域特征,并采用多数表决策略实现两类特征的决策级融合。Tan 等人[23]提出的 IDSFL 网络通过多通道特征调制器和特定情感感知模块捕获整体通道-空间特征,并采用全局平均池化操作提取区域特征。在整体语义提取方面,以上方法仅提取通道-空间单一尺度信息或仅从通道维度提取多尺度特征,忽略了空间多尺度语义的捕获;在区域语义提取方面,以上方法采用单一通道-空间的注意机制或单池化操作,易丢失局部显著性语义,如何捕获通道-空间整体多尺度语义及区域注意语义成为本文亟待解决的问题。此外,以上方法多采用特征级或决策级融合策略,忽略了整体特征与区域特征的相关性,降低了网络在自然场景下的鲁棒性,如何从模型级实现整体特征与区域特征的融合也是本文亟待解决的问题。

为了解决上述问题,本文提出一种通道-空间多尺度增强与双池化注意(channel-spatial multi-scale enhancement and dual-pooling attention,CS-MEDA)的表情识别网络。CS-MEDA 网络由基于通道-空间多尺度结构的整体特征增强子网(whole feature enhancement subnet,WFE)、基于高效通道-空间注意力机制的区域特征注意子网(regional feature attention subnet,RFA)以及基于整体特征引导的特征融合子网(feature fusion subnet,FFS)组成,该网络结构如图 1 所示。在 WFE 中,设计通道对称级联多尺度模块和空间多尺度特征提取模块,以此构建通道-空间多尺度结构获取通道和空间两个维度的整体多尺度语义;在 RFA 中,设计通道注意力模块和空间注意力模块,并将高效局部注意力机制[24]改进为高效通道-空间注意力机制以获取区域的通道和空间显著语义;在 FFS 中,嵌入采用交叉注意力机制进行整体特征与区域特征之间的特征交互,并采用模型级融合策略实现表情识别。

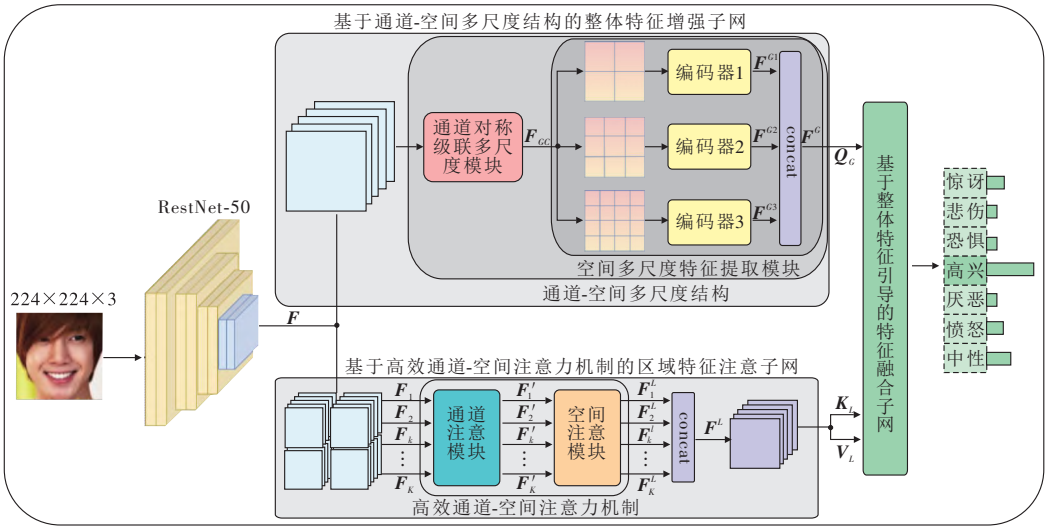

图 1　CS-MEDA 网络结构

Fig. 1　Structure of CS-MEDA

本文主要贡献如下:

a)针对自然场景下整体特征缺乏空间多尺度信息及区域特征易丢失显著语义的问题,提出一种 CS-MEDA 的表情识别网络。与相关方法相比,提出的方法从通道和空间两个维度捕获不同尺度的整体特征以及两种池化策略的区域特征,并采用交叉注意力机制进行整体特征与区域特征的特征交互,从而提升自然场景下的表情识别效果。

b)克服整体通道-空间单一尺度的信息缺乏问题,构建通道-空间多尺度结构。与仅从通道维度捕获整体特征的方法相比,该结构嵌入通道对称级联多尺度模块扩大通道子集感受野,并采用空间多尺度特征提取模块获取不同尺度空间子块的全局相关性,从而实现整体特征的增强。

c)为抑制区域特征易丢失显著性语义的问题,设计高效通道-空间注意力机制。与高效局部注意力机制相比,改进的机制在平均池化基础上增加最大池化操作以捕获通道显著信息,并从空间维度获取空间显著信息,从而提升区域特征的表征能力。

1　CS-MEDA 网络设计

1.1　整体特征增强子网

为获取通道和空间两个维度全局多尺度语义,构建基于通道-空间多尺度结构(channel-spatial multi-scale structure,CSM)的整体特征增强子网,如图 2 所示。该子网主要包括通道对称级联多尺度模块(channel symmetry cascade multi-scale module,CSC)和空间多尺度特征提取模块(spatial multi-scale feature extraction module,SMF)。

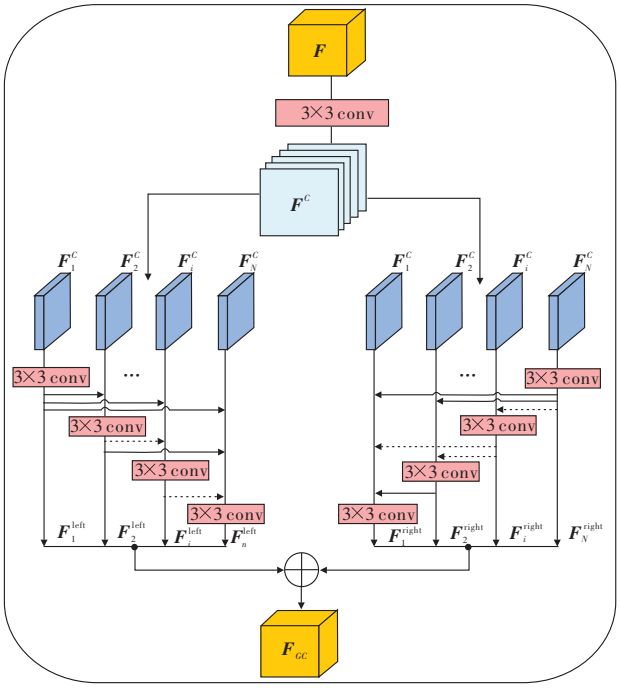

图 2 通道对称级联多尺度模块

Fig. 2 Channel symmetry cascade multi-scale module (CSC)

1.1.1 通道对称级联多尺度模块

首先,以 ResNet50[25] 中间层特征图 $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$($C$ = 1 024、H = 14 和 W = 14 分别表示特征图通道数、长度和宽度)作为 CSC 的输入:

$$\boldsymbol{F}^C = \text{conv}_{3\times 3}(\boldsymbol{F}) \tag{1}$$

其中:$\boldsymbol{F}^C$ 表示经过卷积之后的整体通道特征语义;$\text{conv}_{3\times 3}(\cdot)$ 表示卷积核大小为 3×3 的卷积运算。为获取更精细的通道信息,在通道维度将 $\boldsymbol{F}^C$ 划分成 N 个子集:

$$\boldsymbol{F}_i^C = \text{Split}^C(\boldsymbol{F}^C) \quad 1\leqslant i\leqslant N \tag{2}$$

其中:$\boldsymbol{F}_i^C \in \mathbb{R}^{\frac{C}{N}\times H\times W}$表示第 i 个通道子集;$\text{Split}^C(\cdot)$ 表示通道划分操作。

然后,为了获取通道子集间的交互信息,以对称级联的方式分别从左到右和从右到左方向逐步扩大感受野:

$$\boldsymbol{F}_i^{\text{left}} = \begin{cases} \text{conv}_{3\times 3}^{\text{left}}(\boldsymbol{F}_i^C) & i = 1 \\ \text{conv}_{3\times 3}^{\text{left}}(\boldsymbol{F}_i^C + \sum_1^{i-1}\boldsymbol{F}_i^{\text{left}}) & 1 < i\leqslant N \end{cases} \tag{3}$$

$$\boldsymbol{F}_i^{\text{right}} = \begin{cases} \text{conv}_{3\times 3}^{\text{right}}(\boldsymbol{F}_i^C) & i = N \\ \text{conv}_{3\times 3}^{\text{right}}(\boldsymbol{F}_i^C + \sum_{i+1}^{N}\boldsymbol{F}_i^{\text{right}}) & 1\leqslant i < N \end{cases} \tag{4}$$

其中:$\boldsymbol{F}_i^{\text{left}}$ 和 $\boldsymbol{F}_i^{\text{right}}$ 分别表示第 i 个通道子集对称方向上的多尺度特征。

最后,为表征整体通道间的交互信息,将对称方向上的 N 个多尺度特征进行拼接:

$$\boldsymbol{F}_{GC} = \text{concat}^C(\boldsymbol{F}_1^{\text{left}},\boldsymbol{F}_2^{\text{left}},\cdots,\boldsymbol{F}_i^{\text{left}},\cdots,\boldsymbol{F}_N^{\text{left}}) + \text{concat}^C(\boldsymbol{F}_1^{\text{right}},\boldsymbol{F}_2^{\text{right}},\cdots,\boldsymbol{F}_i^{\text{right}},\cdots,\boldsymbol{F}_N^{\text{right}}) \tag{5}$$

其中:$\text{concat}^C(\cdot)$ 表示通道维度上的特征拼接操作。与单向多尺度相比,设计的 CSC 采用对称结构分别从左到右和从右到左以级联方式捕获各通道子集的多尺度信息,进而提升了整体通道语义的表达能力。

1.1.2 空间多尺度特征提取模块

以上 CSC 关注了通道间的信息交互,但其缺乏空间信息的捕获。为捕获空间维度上的整体信息,传统 ViT[26] 将图像分割为若干个子块,然后采用 Transformer 的多头自注意力机制捕获各子块之间的全局空间依赖关系。然而,该方法仅从单一尺度捕获各子块之间依赖关系,难以抑制自然场景下姿态和遮挡等因素的干扰。因此,为降低姿态和遮挡等因素对整体特征的影响,进一步构建空间多尺度特征提取模块捕获不同空间尺度子块之间的依赖关系,其结构如图 3 所示。

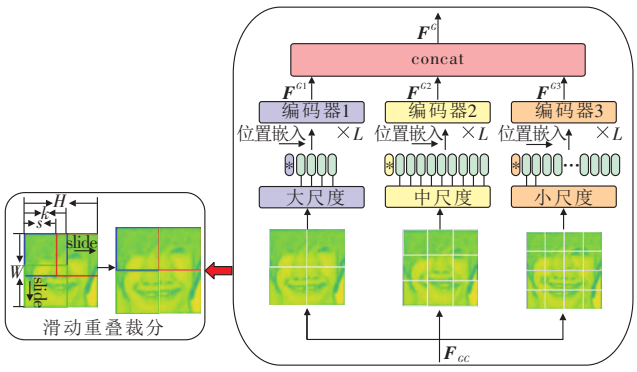

图 3 空间多尺度特征提取模块

Fig. 3 Spatial multi-scale feature extraction module (SMF)

首先,在空间维度上,以 $k\times k$ 为滑动窗口、s 为步幅,将 $W\times H$ 的全局通道多尺度特征 $\boldsymbol{F}_{GC}$ 划分为 N_p 个空间子块 $\{\boldsymbol{F}_i^S\}_{i=1}^{N_p}$:

$$\{\boldsymbol{F}_i^S\}_{i=1}^{N_p} = \text{Split}^S(\boldsymbol{F}_{GC})$$
$$N_p = \left(\frac{W-k}{s}+1\right)\times\left(\frac{H-k}{s}+1\right) \quad 0<s\leqslant k, 0<k\leqslant W \tag{6}$$

其中:$\boldsymbol{F}_i^S$ 表示第 i 个空间子块;$\text{split}^S(\cdot)$ 表示空间划分操作;N_p 表示全局空间划分数。

然后,为适配 Transformer 编码器处理以及保留空间子块的位置信息,将 N_p 个空间子块 $\{\boldsymbol{F}_i^S\}_{i=1}^{N_p}$ 进行图像块嵌入和位置嵌入:

$$\boldsymbol{Z}_0^G = [\boldsymbol{X}_{\text{cls}};\boldsymbol{X}_1^S,\boldsymbol{X}_2^S,\cdots,\boldsymbol{X}_{N_p}^S] + \text{PosE}(N_p+1,k^2C)$$
$$\boldsymbol{X}_i^S = \text{PatchE}(\boldsymbol{F}_i^S) \tag{7}$$

其中:$\text{PosE}(\cdot)$ 表示嵌入的位置索引;$\boldsymbol{X}_{\text{cls}}$ 表示分类标记;$\text{PatchE}(\cdot)$ 表示图像块嵌入;$\boldsymbol{Z}_0^G \in \mathbb{R}^{(N_p+1)\times(k^2C)}$ 表示位置嵌入后的特征序列。

最后,为了获取多层 Transformer 编码器位置信息,将 $\boldsymbol{Z}_0^G$ 输入至 L 层 Transformer 编码器:

$$\hat{\boldsymbol{Z}}_l^G = \text{MHSA}(\text{LN}(\boldsymbol{Z}_{l-1}^G)) + \boldsymbol{Z}_{l-1}^G$$
$$\boldsymbol{Z}_l^G = \text{MLP}(\text{LN}(\hat{\boldsymbol{Z}}_l^G)) + \hat{\boldsymbol{Z}}_l^G \quad 1\leqslant l\leqslant L \tag{8}$$

其中:$\hat{\boldsymbol{Z}}_l^G$ 表示第 l 个编码器中多头注意力机制的输出;$\boldsymbol{Z}_l^G$ 表示第 l 个编码器的输出;L 表示编码器的数量;$\text{LN}(\cdot)$、$\text{MHSA}(\cdot)$ 和 $\text{MLP}(\cdot)$ 分别表示层归一化、多头自注意力机制以及多层感知机。

此外,为了捕获不同尺度子块之间的全局空间依赖关系,本文设计大(k_1 = 8,s_1 = 6)、中(k_2 = 10,s_2 = 2)、小(k_3 = 8,s_3 = 2)三种策略分别提取多尺度特征,并在空间维度上进行拼接:

$$\boldsymbol{F}^G = \text{concat}^S(\boldsymbol{F}^{G1},\boldsymbol{F}^{G2},\boldsymbol{F}^{G3}) \tag{9}$$

其中:$\boldsymbol{F}^G \in \mathbb{R}^{(N_{p1}+N_{p2}+N_{p3}+3)\times(k^2C)}$ 表示整体多尺度增强语义;$\boldsymbol{F}^{G1} \in \mathbb{R}^{(N_{p1}+1)\times(k^2C)}$、$\boldsymbol{F}^{G2} \in \mathbb{R}^{(N_{p2}+1)\times(k^2C)}$、$\boldsymbol{F}^{G3} \in \mathbb{R}^{(N_{p3}+1)\times(k^2C)}$ 分别表示大、中、小三种尺度的整体语义,其中,N_{p1} = 4、N_{p2} = 9、N_{p3} = 16 分别表示三种不同尺度下的整体空间子块数量;$\text{concat}^S(\cdot)$ 表示在空间维度上进行特征拼接。

由上述过程可知,嵌入通道-空间多尺度结构的整体特征增强子网一方面利用 CSC 模块以级联方式从左到右和从右到左方向逐步扩大通道子集感受野范围以捕获通道交互语义,一方面采用 SMF 模块从大、中、小三种尺度获取空间多尺度信息,从而在通道和空间两个维度上实现了整体多尺度语义的增强。其实现步骤如算法 1 所示。

算法 1 整体特征增强子网

输入:$\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ 为 ResNet50 中间层特征图,N 为通道划分数,N_p 为空间划分数,L 为编码器数量。

输出:整体多尺度增强语义。

a) 通道对称级联多尺度模块

$\boldsymbol{F}^C=\mathrm{conv}_{3\times3}(\boldsymbol{F})$

$\{\boldsymbol{F}_i^C\}_{i=1}^N=\mathrm{Split}^C(\boldsymbol{F}^C)$　// 按通道数均分特征

$\boldsymbol{F}_1^{\mathrm{left}}\leftarrow\mathrm{conv}_{3\times3}^{\mathrm{left}}(\boldsymbol{F}_1^C)$

$\boldsymbol{F}_N^{\mathrm{right}}\leftarrow\mathrm{conv}_{3\times3}^{\mathrm{left}}(\boldsymbol{F}_N^C)$

for $i=2$ to N

　$\boldsymbol{F}_i^{\mathrm{left}}\leftarrow\mathrm{conv}_{3\times3}^{\mathrm{left}}(\boldsymbol{F}_i^C+\sum_1^{i-1}\boldsymbol{F}_i^{\mathrm{left}})$ // 左通道多尺度

end for

for $i=1$ to $N-1$

　$\boldsymbol{F}_i^{\mathrm{right}}\leftarrow\mathrm{conv}_{3\times3}^{\mathrm{right}}(\boldsymbol{F}_i^C+\sum_{i+1}^{N}\boldsymbol{F}_i^{\mathrm{right}})$ // 右通道多尺度

end for

$\boldsymbol{F}_{GC}\leftarrow\mathrm{concat}^C(\boldsymbol{F}_1^{\mathrm{left}},\boldsymbol{F}_2^{\mathrm{left}},\cdots,\boldsymbol{F}_i^{\mathrm{left}},\cdots,\boldsymbol{F}_N^{\mathrm{left}})+\mathrm{concat}^C(\boldsymbol{F}_1^{\mathrm{right}},\boldsymbol{F}_2^{\mathrm{right}},\cdots,\boldsymbol{F}_i^{\mathrm{right}},\cdots,\boldsymbol{F}_N^{\mathrm{right}})$

b) 空间多尺度特征提取模块

$\{\boldsymbol{F}_i^S\}_{i=1}^{N_p}=\mathrm{Split}^S(\boldsymbol{F}_{GC})$ // 按空间块划分

$N_p\leftarrow(\frac{W-k}{s}+1)\times(\frac{H-k}{s}+1)\quad 0<s\leqslant k,0<k\leqslant W$

for N_p in {4,9,16} // 大、中、小尺度初始化

　for $i=1$ to N_p//遍历不同空间划分参数

　　$\boldsymbol{X}_i^S\leftarrow\mathrm{PatchE}(\boldsymbol{F}_i^S)$

　　$\boldsymbol{Z}_0^G\leftarrow[\boldsymbol{X}_{\mathrm{cls}};\boldsymbol{X}_1^S,\boldsymbol{X}_2^S,\cdots,\boldsymbol{X}_{N_p}^S]+\mathrm{PosE}(N_p+1,k^2C)$

　end for

　for $l=1$ to L

　　$\hat{\boldsymbol{Z}}_l^G\leftarrow\mathrm{MHSA}(\mathrm{LN}(\boldsymbol{Z}_{l-1}^G))+\boldsymbol{Z}_{l-1}^G$// 多头自注意力

　　$\boldsymbol{Z}_l^G\leftarrow\mathrm{MLP}(\mathrm{LN}(\hat{\boldsymbol{Z}}_l^G))+\hat{\boldsymbol{Z}}_l^G$// 多层感知机

　end for

end for

$\boldsymbol{F}^G\leftarrow\mathrm{concat}^S(\boldsymbol{F}^{G1},\boldsymbol{F}^{G2},\boldsymbol{F}^{G3})$ // 融合不同空间尺度结果

1.2　区域特征注意子网

WFE 提取的整体多尺度增强语义 $\boldsymbol{F}^G$ 包含了面部的全局信息,但其忽略对面部显著情感区域的关注。为了有效捕获区域显著情感信息,CBAM 结构[18]融合通道注意力和空间注意力机制以提取局部特征,但其难以捕获远距离元素之间的依赖关系;高效局部注意力(efficient local attention,ELA)机制[24]采用平均池化(average pooling,AP)操作获取通道位置信息,但其缺乏对显著目标特征和空间信息的关注。为进一步聚焦于人脸表情识别感兴趣区域及突出重要区域的作用,本文在 ELA 基础上引入最大池化(max pooling,MP)操作和空间注意力机制,提出一种基于双池化操作的高效通道-空间注意力(efficient channel-spatial attention mechanism,ECSA)机制。其主要包括通道注意力模块和空间注意力模块,如图 4 所示。

首先,与整体特征增强子网将特征图 $\boldsymbol{F}$ 在通道维度上划分成 N 个子集不同,通道注意力模块将特征图 $\boldsymbol{F}$ 在空间维度划分为 K 个互不重叠的区域子块:

$$\{\boldsymbol{F}_k\}_{k=1}^K=\mathrm{Split}^S(\boldsymbol{F})\tag{10}$$

其中:$\boldsymbol{F}_k$ 表示空间划分后的第 k 个区域子块特征;K 为空间划分块数,参考 MA-Net[16]策略设置为 4。为了获得细粒度的通道特征,在垂直和水平两个空间方向上对每个通道进行双池化操作:

$$\begin{aligned}&\boldsymbol{A}_k^w=\frac{1}{H}\sum_{0\leqslant i<H}\boldsymbol{F}_k(i,w)\quad(0\leqslant w<W)\\&\boldsymbol{M}_k^w=\max_{0\leqslant i<H}\{\boldsymbol{F}_k(i,w)\}\\&\qquad\qquad(1\leqslant k\leqslant K)\\&\boldsymbol{A}_k^h=\frac{1}{W}\sum_{0\leqslant j<W}\boldsymbol{F}_k(h,j)\quad(0\leqslant h<H)\\&\boldsymbol{M}_k^h=\max_{0\leqslant j<W}\{\boldsymbol{F}_k(h,j)\}\end{aligned}\tag{11}$$

其中:$\boldsymbol{F}_k(i,w)$、$\boldsymbol{F}_k(h,j)$ 分别表示区域子块特征 $\boldsymbol{F}_k$ 沿垂直和水平方向的空间信息;$\boldsymbol{A}_k^w$、$\boldsymbol{A}_k^h$ 分别表示用 AP 捕获的垂直和水平方向的一维通道向量;$\boldsymbol{M}_k^w$、$\boldsymbol{M}_k^h$ 分别表示用 MP 捕获的垂直和水平方向的一维通道向量。

然后,采用一维卷积和组规范化操作获取不同方向的位置注意力权重:

$$\begin{aligned}&\boldsymbol{P}_k^{Aw}=\sigma(G_n(F_w(\boldsymbol{A}_k^w))),\ \boldsymbol{P}_k^{Mw}=\sigma(G_n(F_w(\boldsymbol{M}_k^w)))\\&\boldsymbol{P}_k^{Ah}=\sigma(G_n(F_h(\boldsymbol{A}_k^h))),\ \boldsymbol{P}_k^{Mh}=\sigma(G_n(F_h(\boldsymbol{M}_k^h)))\end{aligned}\tag{12}$$

其中:F_w 和 F_h 表示一维卷积;G_n 表示组规范化;σ 表示 sigmoid 函数;$\boldsymbol{P}_k^{Aw}$、$\boldsymbol{P}_k^{Ah}$ 分别表示 AP 捕获的垂直和水平方向上的位置注意力权重;$\boldsymbol{P}_k^{Mw}$、$\boldsymbol{P}_k^{Mh}$ 分别表示用 MP 捕获的垂直和水平方向上的位置注意力权重。获得不同池化操作垂直方向和水平方向的注意力权重后,将双池化操作的输出特征向量分别进行加权融合:

$$\boldsymbol{F}_k^A=\boldsymbol{P}_k^{Ah}\otimes\boldsymbol{P}_k^{Aw}\otimes\boldsymbol{F}_k,\quad\boldsymbol{F}_k^M=\boldsymbol{P}_k^{Mh}\otimes\boldsymbol{P}_k^{Mw}\otimes\boldsymbol{F}_k\tag{13}$$

其中:$\otimes$表示元素级乘法。此外,为进一步获取通道注意显著信息,将 AP 和 MP 操作并行获取的注意特征 $\boldsymbol{F}_k^A$ 和注意特征 $\boldsymbol{F}_k^M$ 进行融合:

$$\boldsymbol{F'}_k=\boldsymbol{F}_k^A+\boldsymbol{F}_k^M\tag{14}$$

其中:$\boldsymbol{F'}_k$ 表示通道注意特征。然而,以上通道注意力模块仅捕获了通道显著信息,但缺乏对局部空间信息的关注。为进一步增强区域空间表征能力,将通道注意特征 $\boldsymbol{F'}_k$ 输入至空间注意力模块,并基于双池化操作获取高效通道-空间注意力特征:

$$\begin{aligned}&\boldsymbol{F}_k^L=\boldsymbol{S}_k\otimes\boldsymbol{F'}_k\\&\boldsymbol{S}_k=\sigma(\mathrm{conv}_{7\times7}(\mathrm{concat}^C(\mathrm{MP}(\boldsymbol{F'}_k),\mathrm{AP}(\boldsymbol{F'}_k))))\end{aligned}\tag{15}$$

其中:$\boldsymbol{F}_k^L$ 表示高效通道-空间注意力特征;$\boldsymbol{S}_k$ 表示空间注意权重;$\mathrm{conv}_{7\times7}(\cdot)$表示卷积核大小为 7×7 的卷积运算。

最后,为描述特征图 $\boldsymbol{F}$ 的局部显著信息,将 K 个区域子块的高效通道-空间注意力特征在空间维度上进行拼接:

$$\boldsymbol{F}^L=\mathrm{concat}^S(\boldsymbol{F}_1^L,\boldsymbol{F}_2^L,\cdots,\boldsymbol{F}_k^L,\cdots,\boldsymbol{F}_K^L)\tag{16}$$

其中:$\boldsymbol{F}^L$ 表示区域双池化显著语义。

由上述过程可知,改进的 ECSA 机制采用双池化操作从通道维度获取通道位置注意信息,并从空间维度获取空间注意信息,从而提升了网络对区域双池化显著语义的提取能力。

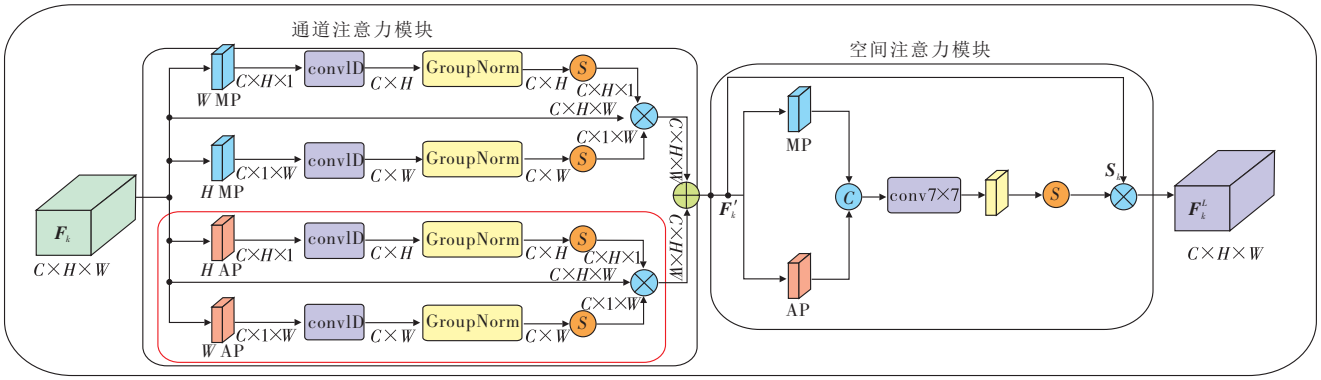

图 4　高效通道-空间注意力机制

Fig. 4　Efficient channel-spatial attention mechanism (ECSA)

1.3 特征融合子网

整体多尺度增强语义和区域双池化显著语义分别从通道和空间两个维度捕获了特征图的全局多尺度信息以及局部显著信息。为充分利用两类特征的互补性,本文采用模型级融合策略构建基于整体特征引导的特征融合子网,如图 5 所示。

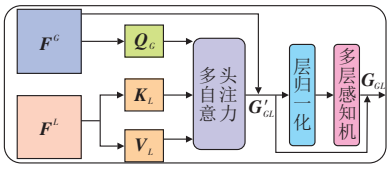

图 5 特征融合子网

Fig. 5 Feature fusion subnet (FFS)

首先,将整体多尺度增强语义 $\boldsymbol{F}^G$ 作为查询向量 $\boldsymbol{Q}_G$,并将区域双池化显著语义 $\boldsymbol{F}^L$ 作为键向量 $\boldsymbol{K}_L$ 和值向量 $\boldsymbol{V}_L$:

$$\boldsymbol{Q}_G = \boldsymbol{F}^G \times \boldsymbol{W}_G^Q, \boldsymbol{K}_L = \boldsymbol{F}^L \times \boldsymbol{W}_L^K, \boldsymbol{V}_L = \boldsymbol{F}^L \times \boldsymbol{W}_L^V \tag{17}$$

其中:$\boldsymbol{Q}_G \in \mathbb{R}^{1\times d}$ 为查询向量;$\boldsymbol{K}_L \in \mathbb{R}^{n\times d}$ 和 $\boldsymbol{V}_L \in R^{n\times d}$ 分别表示键向量和值向量;$\boldsymbol{W}_G^Q \in \mathbb{R}^{d\times d}$、$\boldsymbol{W}_L^K \in \mathbb{R}^{d\times d}$ 和 $\boldsymbol{W}_L^V \in \mathbb{R}^{d\times d}$ 为可训练的权重矩阵。

然后,采用交叉注意力机制和残差结构获取整体与区域特征的相关性:

$$\begin{aligned} \boldsymbol{G}_{GL} &= \mathrm{MLP}(\mathrm{LN}(\boldsymbol{G'}_{GL})) + \boldsymbol{G'}_{GL} \\ \boldsymbol{G'}_{GL} &= \mathrm{softmax}\left(\frac{\boldsymbol{Q}_G \boldsymbol{K}_L^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_L + \boldsymbol{F}_G \end{aligned} \tag{18}$$

其中:$\boldsymbol{G}_{GL}$ 表示多头注意力的融合特征;softmax 为归一化指数函数。

最后,将融合特征 $\boldsymbol{G}_{GL}$ 输入 GAP 层、FC 层以及 softmax 分类器。获得类别向量 $\boldsymbol{O} = (o_1, \cdots, o_u, \cdots, o_U)$ 及表情类别 c:

$$\boldsymbol{O} = \mathrm{softmax}(\mathrm{FC}(\mathrm{GAP}(\boldsymbol{G}_{GL}))) \tag{19}$$

$$P(o_u) = \frac{e^{o_u}}{\sum_{v=1}^{U} e^{o_v}} \quad u \in [1, U] \tag{20}$$

$$c = \mathrm{argmax}\{P(o_u)\}_{u=1}^{U} \tag{21}$$

其中:U 表示表情类别数;$P(o_u)$ 为融合特征 $\boldsymbol{G}_{GL}$ 被分类为第 u 类表情的概率。

2 CS-MEDA 网络训练

为优化 CS-MEDA 网络,本文采用子网独立优化与双子网协同优化相结合的策略。在子网独立优化过程中,WFE 和 RFA 的目标函数 $loss_G$ 和 $loss_L$ 均采用交叉熵损失:

$$\begin{aligned} loss_G &= -\sum_{u=1}^{U} y_u \ln(P(a_u)) \\ loss_L &= -\sum_{u=1}^{U} y_u \ln(P(b_u)) \end{aligned} \tag{22}$$

其中:$\boldsymbol{a} = (a_1, a_2, \cdots, a_u, \cdots, a_U)$ 和 $\boldsymbol{b} = (b_1, b_2, \cdots, b_u, \cdots, b_U)$ 分别为 WFE 和 RFA 的类别向量;$\boldsymbol{y} = (y_1, y_2, \cdots, y_u, \cdots, y_U)$ 为真实表情的标签向量。

在双子网协同优化过程中,以 WFE 和 RFA 独立优化的最优权重为 CS-MEDA 协同优化的初始值,并构建整体网络的目标优化函数 $loss$:

$$loss = \alpha loss_G + (1-\alpha) loss_L + loss_F \tag{23}$$

其中:$loss$ 为 CS-MEDA 网络的优化函数;$loss_F = -\sum_{u=1}^{U} y_u \ln(P(o_U))$ 表示融合特征的损失函数;$\alpha \in [0,1]$ 为 $loss_G$ 和 $loss_L$ 的协同优化权重,随着 α 增大,WFE 的影响逐渐增强;反之,RFA 的影响逐渐增强。

3 实验设计与分析

3.1 数据集和实验环境

为了说明 CS-MEDA 网络的有效性,本文在 RAF-DB[27] 和 FERPlus[28] 数据集上进行实验。RAF-DB[27] 数据集包含由 Flickr API 收集的15 339张基本表情图像,共包括 7 种基本表情(高兴、惊讶、悲伤、愤怒、厌恶、恐惧和中性)。在 RAF-DB[27] 数据集上,本文随机选取12 271张图像用于训练,其余的用于测试。与 RAF-DB[27] 数据集不同,FERPlus[28] 数据集除了包含 7 种基本表情,增加了蔑视表情的标注。在 FERPlus[28] 数据集上,本文随机选取28 709张和3 589张图像分别用于训练和测试。此外,为了进一步验证 CS-MEDA 网络对姿态变化及遮挡等因素的影响,分别在 RAF-DB-Pose > 30°、RAF-DB-Pose > 45°、RAF-DB-Occlusion、FERPlus-Pose > 30°、FERPlus-Pose > 45°、FERPlus-Occlusion 等数据集上进行实验验证。此外,本文在 Intel® Xeon® Platinum 8255C CPU @ 2.50 GHz、RTX 2080Ti GPU 等计算平台上,基于 PyTorch 深度学习框架实现了 CS-MEDA网络的构建和训练,具体参数如表 1 所示。

表 1 CS-MEDA 网络参数设置

Tab. 1 Parameter settings of CS-MEDA network

参数名称	参数设置
输入图像大小	224 × 224
训练学习率	0.01
衰减系数	0.95
优化器	随机梯度下降
迭代次数	70
动量	0.9
全局通道划分数 N	4
网络批量	32
大、中和小尺度全局空间划分数 Np_1、Np_2 和 Np_3	4、9、16
区域划分块数 K	4
优化器权重	0.000 1
协同优化权重 α	0.6

3.2 实验分析

3.2.1 不同数据集上的识别性能

为了说明 CS-MEDA 网络在不同人脸表情数据集上的识别性能,在 8 种自然场景数据集上分别统计了各类表情的识别率及平均识别效果,如表 2(“—”表示无蔑视表情)所示。由表 2 可知,其在 RAF-DB 和 FERPLus 数据集上的平均识别率分别为 89.97% 和 90.26%,其中“高兴”在 CS-MEDA 网络上表现最佳,表情识别率均超 95%。这表明提出的网络能够充分利用整体多尺度增强语义和区域显著语义的互补优势。与 RAF-DB和 FERPlus 原始数据集相比,两者在遮挡数据集的表情识别率方面有所下降,但仍维持较高水平。这表明尽管面部区域遮挡易致使通道-空间多尺度结构难以捕获面部整体多尺度增强语义,但改进的 ECSA 机制能够通过双池化操作有效捕获区域显著语义,从而提升遮挡数据集上的表情识别效果。在 RAF-DB 和 FERPlus 的姿态变化数据集上,Pose > 45° 的识别率略低于 Pose > 30°,这说明 CS-MEDA 网络一方面通过多尺度特征结构和双池化操作分别获取通道-空间整体增强语义和区域显著语义,另一方面采用交叉注意力机制实现整体多尺度增强语义对区域双池化显著语义的引导,以捕获长距离依赖关系,弥补了姿态变化导致的面部信息不完整的问题,从而提升了 CS-MEDA网络的表情识别性能。

为了说明提出的网络在不同数据集上各类别识别性能,分别统计了 CS-MEDA 网络在 RAF-DB、FERPlus 及其遮挡与姿态变化数据集上的混淆矩阵,如图 6 所示。由图 6 可知,在 RAF-DB 与姿态变化数据集上,除“厌恶”外,其他类别的表情识别率均超过 85%;在 FERPlus 与姿态变化数据集上,除“蔑视”外,其他类别的表情识别率均超过 80%。同时,易混淆的表情类别集中于“厌恶”与“悲伤”以及“蔑视”与“中性”等情感类别上。这主要因为“厌恶”“悲伤”“蔑视”等负向面部表情的

呈现形式在客观上存在较高的相似性;此外,人类情感类别标注的主观不一致性也加剧了这些表情类别的区分难度。与IDSFL网络混淆矩阵[23]相比,CS-MEDA网络不仅在易于识别的情感类别(如“高兴”“惊讶”等)上取得了与其相近的识别率,而且较大幅度提升了难以识别情感类别的识别效果,如在“厌恶”“悲伤”等情感类别上的识别率提升均超过了10%。此外,IDSFL网络在“厌恶”“悲伤”等情感类别的样本上最高误识别率达到18%,而本文网络在相关类别的误识别率均在10%以下。以上结果表明:提出的CS-MEDA网具有较高类别识别率且能够维持较好的类别均衡性。

此外,为了说明CS-MEDA网络的运算效率,本文统计了ResNet50[25]、WFE、RFA、FFS以及CS-MEDA的参数量、FLOPs、训练时间和测试时间,如表3所示。在表3中,相较于ResNet50[25],CS-MEDA网络的参数量显著上升。但得益于本文设计的协同优化方案,网络的FLOPs以及在两个数据集上的训练耗时与测试耗时并未显著增加。同时,在RAF-DB和FERPlus数据集上,CS-MEDA网络分别识别3 068张和3 589张测试集图片时间为32.11 s和36.12 s,平均单张图片耗时10.5 ms和10.1 ms。从单张图片的处理时间来看,提出的CS-MEDA能够达到相关场景的实时性要求。

表2　CS-MEDA网络在不同人脸表情数据集的识别性能

Tab.2　Recognition performance of CS-MEDA network on different facial expression datasets

数据集	高兴	悲伤	恐惧	愤怒	惊讶	厌恶	中性	蔑视	平均识别率/%
RAF-DB	96.67	90.67	89.16	86.01	91.23	81.11	88.91	—	89.97
Occlusion-RAF-DB	95.87	83.20	84.51	85.52	84.97	84.98	80.31	—	85.62
Pose>30°-RAF-DB	97.12	92.35	86.89	86.52	90.56	87.72	89.68	—	90.12
Pose>45°-RAF-DB	96.17	92.82	86.08	87.16	90.71	85.22	89.51	—	89.67
FERPlus	96.31	87.79	88.35	90.59	91.78	92.69	90.56	84.02	90.26
Occlusion-FERPlus	95.58	82.01	80.53	87.32	89.67	86.75	89.01	78.43	86.16
Pose>30°-FERPlus	96.28	88.05	82.73	90.32	92.67	93.75	91.08	79.83	89.34
Pose>45°-FERPlus	96.46	88.13	81.35	91.49	93.44	92.98	90.20	79.31	89.17

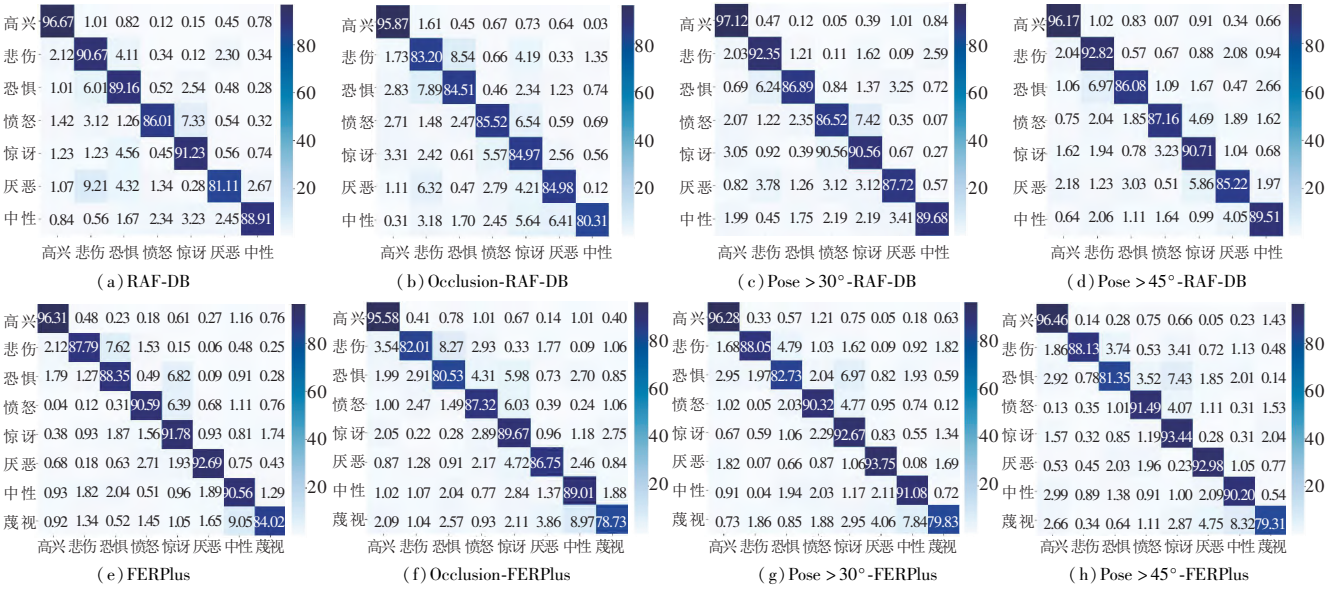

图6　CS-MEDA网络在不同人脸表情数据集的混淆矩阵

Fig.6　Confusion matrix of CS-MEDA network on different facial expression datasets

表3　CS-MEDA网络的模型复杂度及运算效率

Tab.3　Model complexity and computational efficiency of CS-MEDA network

网络	参数量/M	FLOPs/G	训练时间/h		测试时间/s	
			RAF-DB	FERPlus	RAF-DB	FERPlus
ResNet50	23.52	4.12	1.03	2.42	13.33	15.08
WFE	38.52	5.13	1.18	1.53	26.10	28.67
RFA	25.12	1.33	0.76	1.32	11.34	12.34
FFS	12.14	0.12	—	—	—	—
CS-MEDA(本文网络)	89.56	6.85	1.41	2.62	32.11	36.12

3.2.2　超参数对表情识别性能的影响

1)通道划分数N对网络性能的影响

为了分析CSC中通道划分数对CS-MEDA网络性能的影响,在两个数据集上分别统计5种通道划分策略的平均表情识别率,如图7所示。由图7可知,当$N<4$时,随着通道划分数量的上升,提出的对称级联多尺度模块能够逐层扩大通道子集的感受野,因此其表情识别率逐渐提高;当$N=4$时,在两个数据集上,其表情识别率达到最佳,分别为89.97%和90.26%;当$N>4$时,随着分组数量继续增大,通道子集的粒度信息逐渐减少,故表情识别率呈下降趋势。

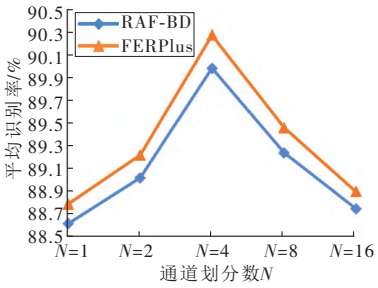

图7　通道划分数N对网络性能的影响

Fig.7　Influence of channel division number N on network performance

2)空间划分数N_p对网络性能的影响

为了说明全局空间划分数N_p对CS-MEDA网络识别率的

影响,本文设计 11 种划分策略,如表 4 所示。其中,K1 表示无空间划分操作策略;K2 ~ K4 表示 $N_p=4$ 时仅包含的 3 种不同滑动窗口及步幅取值策略;K5 和 K6 表示 $N_p=9$ 时仅包含的 2 种不同滑动窗口及步幅取值策略;K7、K8、K9 和 K10 分别表示 N_p 为 16($k=8,s=2$)、25($k=6,s=2$)、36($k=4,s=2$)、49($k=2,s=2$)时的唯一划分策略。从表 4 可以看出,与无空间划分操作策略相比,采用空间划分策略的表情识别性能均有所提升,这表明空间划分有助于捕获各子块之间的全局空间依赖关系,从而更有效地提取全局空间信息。随着空间划分数 N_p 的增大,CS-MEDA 网络捕获的全局空间上下文信息逐渐增强,因此其识别性能得以提升;当 $N_p=9$ 时,网络的识别性能达到最佳;然而,随着 N_p 进一步增大,网络的表情识别率呈下降趋势。这表明过多数量的空间子块将易导致相邻块间的语义关联不完整,从而难以有效捕获全局空间语义。此外,在大、中、小尺度划分策略中,策略 K4($k=8,s=6,N_{p1}=4$)、策略 K5($k=10,s=2,N_{p2}=9$)以及策略 K7($k=8,s=2,N_{p3}=16$)的识别性能表现最佳。

表 4 空间划分数 N_p 对网络性能的影响

Tab. 4 Influence of spatial division number N_p on network performance

划分策略	滑动窗口大小/k	步幅 s	空间子块分大小	划分尺度	空间划分数 N_p	RAF-DB /%	FERPlus /%
K1	—	—	14×14	无划分	1	87.57	87.64
K2	12	2	12×12			87.76	87.91
K3	10	4	10×10	大尺度	4	88.12	88.14
K4	8	6	8×8			88.45	88.56
K5	10	2	10×10	中尺度	9	89.04	89.17
K6	6	4	6×6			88.89	89.12
K7	8	2	8×8		16	88.76	89.05
K8	6	2	6×6	小尺度	25	88.71	89.01
K9	4	2	4×4		36	88.56	89.01
K10	2	2	2×2		49	88.34	88.89

为了进一步说明大、中、小尺度相互组合对 CS-MEDA 网络识别率的影响,本文引入 7 种组合策略,如表 5 所示。在表 5 中,相较于 S1 ~ S3 的单尺度策略,S4 ~ S6 的双尺度策略在 CS-MEDA 网络上表情识别率均有提升,这说明大、中、小三种尺度两两组合的效果优于仅包含单尺度的效果。相较于双尺度策略,S7 的多尺度策略(本文网络)具有更好的表情识别效果,在 RAF-DB 和 FERPlus 数据集上的识别率分别达到 89.97% 和 90.26%。这说明设计的 SMF 能够从空间维度捕获整体多尺度增强信息,从而提升 CS-MEDA 网络的识别性能。

表 5 不同尺度组合策略对网络性能的影响

Tab. 5 Influence of different scale combination strategies on network performance

组合策略	N_{p1} ($k=8,s=6$)	N_{p2} ($k=10,s=2$)	N_{p3} ($k=8,s=2$)	RAF-DB /%	FERPlus /%
S1	4	0	0	88.45	88.56
S2	0	9	0	89.04	89.05
S3	0	0	16	88.76	89.17
S4	4	9	0	89.47	89.24
S5	4	0	16	89.25	89.62
S6	0	9	16	89.56	89.78
S7(本文网络)	4	9	16	89.97	90.26

3)协同优化权重 α 对网络性能的影响

为了分析协同优化权重 α 对表情识别性能的影响,本文统计了 $\alpha\in[0,1]$ 时的 CS-MEDA 网络表情识别率,如图 8 所示。图 8 表明:当 $\alpha=0$ 或 $\alpha=1$ 时,由于缺乏 WFE 或 RFA 的补偿信息,CS-MEDA 网络表情识别率较低;当 $0<\alpha<0.6$ 时,WFE 的影响逐渐增强,CS-MEDA 网络表情识别率在 RAF-DB 和 FERPlus 数据集上均呈上升趋势,这表明 WFE 与 RFA 两类子网的协同作用有助于提高该网络的识别性能;当 $\alpha=0.6$ 时,在两个数据集上的表情识别率达到最佳,分别为 89.97% 和 90.26%;当 $0.6<\alpha<1$ 时,RFA 的作用被逐渐忽略,CS-MEDA 网络在两个数据集上的表情识别率呈下降趋势。根据 CS-MEDA网络在两个数据集上的性能表现,本文选择 $\alpha=0.6$。

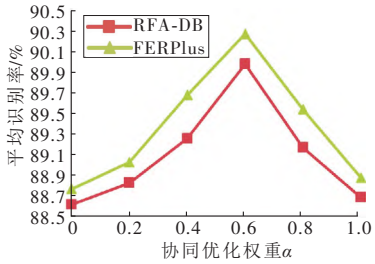

图 8 不同协同优化权重 α 对网络性能的影响

Fig. 8 Influence of different collaborative optimization weights α on network performance

3.2.3 消融实验

为了分析 CS-MEDA 网络中各模块的作用,将 CSC、SMF、ECSA 及 FFS 进行 11 种形式的组合,组合方式及实验结果如表 6 所示。与基线网络的策略 G1(ResNet50[25])相比,策略 G2、G3 和 G4 的表情识别率分别提升了 11.91 和 9.40 个百分点,12.02 和 9.33 个百分点,12.08 和 9.36 个百分点。这表明构建的 CSC、SMF 模块和改进的 ECSA 机制能够分别提升网络的表情识别性能。与 G2、G3 和 G4 仅嵌入单一模块提取整体或区域语义不同,G5 级联 CSC 和 SMF 模块提取通道-空间整体语义且不包含区域语义,G6 和 G7 则分别采用 CSC + ECSA 和 SMF + ECSA 获取通道-空间整体和区域语义。实验结果表明,CSC、SMF 和 ECSA 三种模块两两组合的表情识别效果优于单个模块。与嵌入 CSC 和 SMF 的策略 G5 相比,策略 G6 和 G7 在 WFE 上分别嵌入 CSC 和 SMF,其表情识别率在两个数据集上分别提升了 0.24 和 0.22 个百分点以及 0.15 和 0.17 个百分点。这说明整体-区域语义融合的效果优于仅采用单一整体语义策略。与策略 G5、G6 和 G7 相比,策略 G8 嵌入通道-空间多尺度结构和 ECSA 机制,在两个数据集上表情识别率分别提高了 0.75 和 0.90 个百分点,0.51 和 0.68 个百分点,0.60 和 0.73 个百分点。这表明构建的通道-空间多尺度结构能够捕获整体多尺度增强语义,改进的 ECSA 机制能够获取区域双池化显著语义,从而有利于提升自然场景下的表情识别效果。与策略 G6、G7 和 G8 相比,策略 G9、G10 和 G11 在 FFS 中采用交叉注意力机制实现整体多尺度增强语义和区域双池化显著语义的模型级融合,其表情识别率进一步分别提高了 0.30 和 0.22 个百分点,0.24 和 0.16 个百分点,0.25 和 0.24 个百分点。

表 6 不同消融策略对网络性能的影响

Tab. 6 Influence of different ablation strategies on network performance

消融策略	CSC	SMF	ECSA	FFS	RAF-DB	FERPlus
G1(基线网络)					76.43	79.31
G2	√				88.34	88.71
G3		√			88.45	88.64
G4			√		88.51	88.67
G5	√	√			88.97	89.12
G6	√		√		89.21	89.34
G7		√	√		89.12	89.29
G8	√	√	√		89.72	90.02
G9	√		√	√	89.51	89.56
G10		√	√	√	89.36	89.45
G11(本文网络)	√	√	√	√	89.97	90.26

同时,为了说明设计的通道对称级联多尺度模块在整体特征增强子网中的作用,本文将其与通道单向多尺度模块和通道对称多尺度模块进行比较,如表 7 所示。由表 7 可知:与通道

单向多尺度模块相比,通道对称多尺度模块能够从双向扩大通道子集感受野范围,其在 RAF-DB 和 FERPlus 数据集上的识别率分别提高了 0.56 和 0.61 个百分点。与通道对称多尺度模块相比,本文提出的通道对称级联多尺度模块在两个数据集上的表情识别率分别提高了 0.76 和 0.81 个百分点。这表明 CSC 在对称结构基础上引入通道子集级联方式能够进一步提升表情识别性能。

表 7　不同全局通道多尺度模块对网络性能的影响

Tab. 7　Influence of different global channel multi-scale modules on network performance

全局通道多尺度模块	RAF-DB/%	FERPlus/%
通道单向多尺度模块	88.65	88.84
通道对称多尺度模块	89.21	89.45
通道对称级联多尺度模块(CSC)	89.97	90.26

此外,为了说明改进的 ECSA 在 RFA 中的有效性,本文在 RAF-DB 和 FERPlus 数据集上比较了不同注意力机制下的表情识别效果,如表 8 所示。在表 8 中,与策略 ELA(策略 R1)相比,策略 R2 的表情识别率分别提高 0.41 和 0.51 个百分点;与策略 R4 相比,策略 R5 的表情识别率分别提高了 0.41 和 0.54 个百分点。

表 8　不同局部注意力机制对表情识别性能的影响

Tab. 8　Influence of different local attention mechanisms on expression recognition performance

消融策略	ELA[24]	通道注意力模块	空间注意力模块	RAF-DB/%	FERPlus/%
R1	√			89.37	89.45
R2		√		89.78	89.96
R3			√	89.39	89.55
R4	√		√	89.56	89.72
R5(本文策略)		√	√	89.97	90.26

这说明通道注意力模块在 ELA 单一的 AP 操作基础上增加 MP 操作能够较好地捕获通道显著信息。相较于仅采用通道注意力模块(策略 R2)或空间注意力模块(策略 R3),策略 R5 采用双池化操作获取通道-空间区域显著信息,进一步提升了面部局部细节的捕获能力。

为了直观地描述 CS-MEDA 网络的识别效果,采用 Grad-CAM[29] 分别对 ResNet50、WFE、RFA、MA-Net[16]、SCAN + CCI[20] 以及 CS-MEDA(本文网络)的关注效果进行可视化,如图 9 所示。图 9 热力图表明:由于姿态变化、遮挡等因素的干扰,ResNet50 的关注区域包含了较多的非表情区域;与 ResNet50 相比,WFE 采用通道-空间多尺度结构捕获整体多尺度增强信息,其关注区域的边缘更为清晰;同时,RFA 通过高效通道-空间注意力机制获取区域显著语义,其能够关注眼睛、鼻子、嘴巴等主要表情区域。特别地,为了进一步说明 CS-MEDA 网络在姿态变化数据集上的识别效果,分别对 Pose > 30°-RAF-DB、Pose > 45°-RAF-DB 数据集上七种表情类别的关注效果进行可视化,如图 10 所示。图 10 热力图表明:尽管存在不同程度的姿态偏转,CS-MEDA 网络仍然能够定位到五官区域并聚焦于眼睛和嘴角等局部细节,因此其在姿态变化数据集上保持了较好的识别性能;此外,在负向表情类别上,如“厌恶”“悲伤”等,CS-MEDA 网络也能够较精确地关注到面部五官区域,从而有利于提升此类易混淆表情的识别效果和维持较好的类别均衡性。与 CS-MEDA 网络相比,MA-Net[16] 和 SCAN + CCI[20] 的热力图虽然能够定位五官区域,但其在聚焦眼睛和嘴角等表情变化的细节上略显不足。这表明:CS-MEDA 从通道和空间两个维度捕获不同尺度的整体特征以及不同池化策略的区域特征,并采用整体语义引导的模型级融合策略充分利用不同语义间的互补信息,从而有利于克服姿态偏转、遮挡等因素的影响。

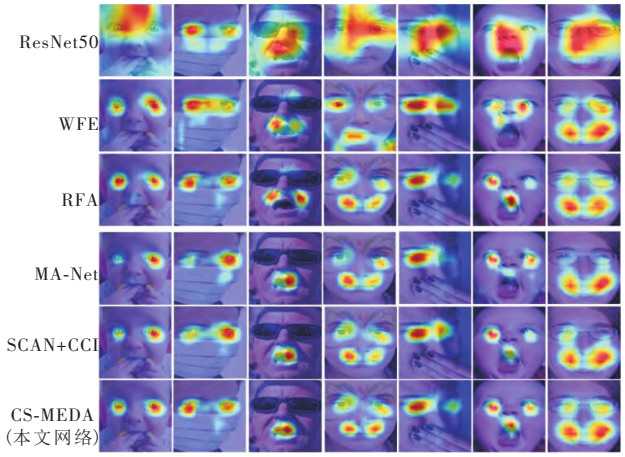

图 9 在 RAF-DB 数据集上的相关网络热力图比较

Fig. 9　Comparison of related network heatmaps on RAF-DB dataset

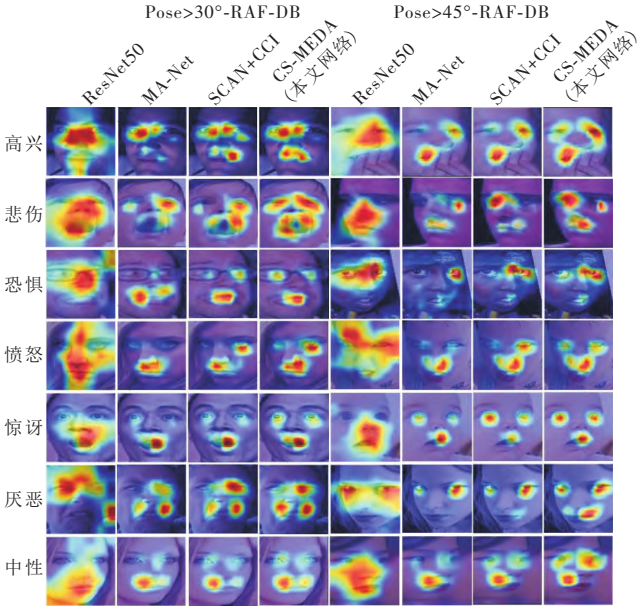

图 10　在 Pose > 30°-RAF-DB、Pose > 45°-RAF-DB 数据集上各表情类别的相关网络热力图比较

Fig. 10　Comparison of related network heatmaps for each expression category on Pose > 30°-RAF-DB and Pose > 45°-RAF-DB datasets

3.2.4　与相关方法比较

在 RAF-DB 和 FERPlus 数据集上,将 CS-MEDA 网络与基于面部整体语义、面部区域语义以及两者语义融合的表情识别方法进行比较,结果如表 9 所示(“–”表示该文献未提供相关数据)。在表 9 中,与基线的 ResNet50[25] 相比,嵌入通道-空间多尺度结构和 ECSA 机制的 CS-MEDA 网络表情识别率分别提升了 13.54 和 10.95 个百分点。在基于面部整体语义[9,10] 和基于面部区域语义[12,13] 中,整体语义的 CVT[9] 和区域语义的 FER-VT[13] 表现最佳,而提出的 CS-MEDA 网络在两个数据集上较以上两者分别提升了 1.83 和 1.45 个百分点以及 1.71 和 0.22 个百分点,这表明整体语义与区域语义融合策略充分发挥两类语义的互补优势,从而提升了 CS-MEDA 网络表情识别性能。与基于整体-区域语义融合方法[15,16,19~21,23,30,31] 相比,提出的 CS-MEDA 网络分别获取通道-空间整体多尺度增强语义以及区域双池化显著语义,并基于交叉注意力实现两类语义的模型级融合,因此获得了最佳的识别效果。

此外,为了说明自然场景下遮挡和姿态变化等因素对相关方法的影响,分别统计了两个数据集不同场景下的表情识别结

果及采用的基线网络,如表 10 所示。在表 10 中,本文网络在各个数据集上均展现出最优的表情识别效果,且与基线的 ResNet50[25] 相比,CS-MEDA 网络表情识别率在各个数据集上提升了 10.07 ~ 13.63 个百分点。与仅捕获单尺度全局空间信息的 CVT[9] 网络相比,CS-MEDA 网络通过通道-空间多尺度特征结构分别从通道和空间两个维度获取整体多尺度增强信息,降低了网络对姿态变化和遮挡的敏感度;与 RAN[12] 和 FER-VT[13] 等利用关系注意力或网格注意力来获取区域特征不同,CS-MEDA 网络采用双池化操作从通道和空间两个维度获取区域显著语义,有效地减少了姿态变化和遮挡的影响。在基于面部整体-区域语义的表情识别方法中,与 OADN[30]、MA-Net[16]、SCAN + CCI[20]、AMP-Net[19] 和 IDSFL[23] 等采用特征级融合或决策级融合相比,CS-MEDA 网络在 RAF-DB 和 FERPlus 的姿态变化和遮挡数据集上的平均识别率均有提升。特别地,与 IDSFL 网络[23] 相比,CS-MEDA 网络在 FERPlus-Pose > 45°数据集上的识别性能提升了 2.61 个百分点。这说明 CS-MEDA 网络一方面通过多尺度结构和双池化操作分别获取通道-空间整体多尺度增强语义和区域双池化显著语义,另一方面采用交叉注意力机制实现整体多尺度增强语义对区域双池化显著语义的引导,从而捕获长距离依赖关系以抑制头部偏转以及面部遮挡的影响。

表 9 不同网络在 RAF-DB 和 FERPlus 数据集上的识别性能

Tab. 9 Recognition performance of different networks on RAF-DB and FERPlus datasets

方法		年份	基线网络	RAF-DB /%	FERPlus /%
基线网络	ResNet50[25]	2016	ResNet50	76.43	79.31
面部整体语义	LBAN-IL[10]	2021	ResNet18	85.89	—
	CVT[9]	2023	ResNet18	88.14	88.81
面部区域语义	RAN[12]	2020	ResNet18	86.90	88.55
	FER-VT[13]	2021	ResNet34	88.26	90.04
面部整体-区域语义融合	OADN[30]	2020	ResNet50	89.83	—
	MA-Net[16]	2021	ResNet18	88.40	—
	SCAN + CCI[20]	2021	ResNet50	89.02	89.42
	AMP-Net[19]	2022	ResNet34	89.25	—
	PACVT[15]	2023	ResNet18	88.21	88.72
	CFNet[21]	2023	—	87.52	—
	EDGL-FLP[31]	2023	ResNet50	89.90	—
	IDSFL[23]	2024	ResNet18	89.54	90.06
	CS-MEDA (本文网络)	2024	ResNet50	89.97	90.26

表 10 不同网络在姿态变化和遮挡数据集上的识别性能

Tab. 10 Recognition performance of different networks on posture variation and occlusion datasets

方法		年份	基线网络	RAF-DB/%			FERPlus/%		
				Occlusion	Pose > 30°	Pose > 45°	Occlusion	Pose > 30°	Pose > 45°
基线网络	ResNet50[25]	2016	ResNet50	74.45	78.64	76.04	75.12	79.27	77.27
面部整体语义	CVT[9]	2023	ResNet18	83.95	87.97	88.35	84.79	88.29	87.20
面部区域语义	RAN[12]	2020	ResNet18	82.72	86.74	85.20	83.63	82.23	80.40
	FER-VT[13]	2021	ResNet34	84.32	88.03	86.08	85.24	88.56	87.06
面部整体-区域语义融合	OADN[30]	2020	ResNet50	—	—	—	84.57	88.52	87.50
	MA-Net[16]	2021	ResNet18	83.65	87.89	87.89	—	—	—
	SCAN + CCI[20]	2021	ResNet50	85.03	89.82	89.07	86.12	88.89	88.15
	AMP-Net[19]	2022	ResNet34	85.28	89.75	89.25	85.44	88.52	87.57
	IDSFL[23]	2024	ResNet18	85.03	89.25	88.35	86.00	88.35	86.56
	CS-MEDA(本文网络)	2024	ResNet50	85.62	90.12	89.67	86.16	89.34	89.17

4 结束语

为克服自然场景下整体特征缺乏空间多尺度信息及局部特征易丢失显著语义的问题,构建一种 CS-MEDA 网络的表情识别框架。在 WFE 中,提出多尺度结构分别从通道和空间维度捕获不同尺度的通道交互信息和空间依赖关系,实现整体多尺度特征的增强;在 RFA 中,鉴于单一池化操作的 ELA 机制缺乏对显著目标特征和空间信息的关注,设计基于双池化操作 ECSA 机制,获取通道-空间区域显著语义;在 FFS 中,采用交叉注意力机制实现整体多尺度增强语义对区域双池化显著语义引导,完成两类特征的模型级融合。实验分析了 CS-MEDA 网络在自然场景数据集 RAF-DB 和 FERPlus 上的表情识别性能,分别比较了 5 种全局通道划分策略、7 种全局空间划分组合策略以及 11 种消融实验策略对表情识别率的影响;此外,对 CS-MEDA 网络三个子网的热力图进行了具体分析和解释。相关结果表明:提出的网络在两自然场景数据集上具有 89.97% 和 90.26% 的表情识别率,较基线网络分别提升了 13.54 和 10.95 个百分点,且在姿态变化、遮挡等数据集上具有较好的识别性能。然而,在高效通道-空间注意力机制中,双池化操作虽然提升了局部特征的提取能力,但降低了特征图的分辨率,限制了对细节的全面捕捉。为了有效缓解了由于池化带来的空间信息丢失问题,可借助深度可分离卷积在保留空间细节方面的优势,从而在下采样过程中保留特征图更多的细粒度信息。此外,由于受面部遮挡、姿态偏转、光照强度、样本数量、标注噪声等因素的影响,CS-MEDA 网络的表情识别性能仍有进一步提升的空间。如何借助大模型在特征提取、语义理解等方面的优势及采用预训练与迁移学习相结合的方式增强网络的抗干扰能力,将是未来的主要研究工作;同时,当前提出的 CS-MEDA 网络主要适用于自然场景下单帧图像的情感识别,如何引入时序信息实现动态人脸表情识别也将是下一步着力解决的问题。

参考文献:

[1] Saadi I, Cunningham D W, Taleb-Ahmed A, *et al*. Driver's facial expression recognition: a comprehensive survey [J]. Expert Systems with Applications, 2024, 242: article ID 122784.

[2] 蒋斌, 崔晓梅, 江宏彬, 等. 轻量级网络在人脸表情识别上的新进展 [J]. 计算机应用研究, 2024, 41(3): 663-670. (Jiang Bin, Cui Xiaomei, Jiang Hongbin, *et al*. New advances in lightweight networks for facial expression recognition [J]. Application Research of Computers, 2024, 41(3): 663-670.)

[3] 张为, 李璞. 基于注意力机制的人脸表情识别网络 [J]. 天津大学学报: 自然科学与工程技术版, 2022, 55(7): 706-713. (Zhang Wei, Li Pu. Facial expression recognition network based on attention mechanism [J]. Journal of Tianjin University: Science and Technology, 2022, 55(7): 706-713.)

[4] Sajjad M, Ullah F U M, Ullah M, *et al*. A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines [J]. Alexandria Engineering Journal, 2023, 68: 817-840.

[5] 胡敏, 胡鹏远, 葛鹏, 等. 基于面部运动单元和时序注意力的视频表情识别方法 [J]. 计算机辅助设计与图形学学报, 2023, 35(1): 108-117. (Hu Min, Hu Pengyuan, Ge Peng, *et al*. Video ex-

pression recognition method based on facial motion unit and temporal attention [J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(1): 108-117.)

[6] 陈公冠, 张帆, 王桦, 等. 区域增强型注意力网络下的人脸表情识别 [J]. 计算机辅助设计与图形学学报, 2024, 36(1): 152-160. (Chen Gongguan, Zhang Fan, Wang Hua, *et al*. Facial expression recognition based on region enhanced attention network [J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(1): 152-160.)

[7] Xu Rui, Huang Aibin, Hu Yuanjing, *et al*. GFFT: global-local feature fusion transformers for facial expression recognition in the wild [J]. Image and Vision Computing, 2023, 139: article ID 104824.

[8] 何昱均, 韩永国, 张红英. FFDNet: 复杂环境中的细粒度面部表情识别 [J]. 计算机应用研究, 2024, 41(5): 1578-1584. (He Yujun, Han Yongguo, Zhang Hongying. FFDNet: fine-grained facial expression recognition in challenging environments [J]. Application Research of Computers, 2024, 41(5): 1578-1584.)

[9] Ma Fuyan, Sun Bin, Li Shutao. Facial expression recognition with visual transformers and attentional selective fusion [J]. IEEE Trans on Affective Computing, 14(2): 1236-1248.

[10] Li Hangyu, Wang Nannan, Yu Yi, *et al*. LBAN-IL: a novel method of high discriminative representation for facial expression recognition [J]. Neurocomputing, 2021, 432: 159-169.

[11] Kopalidis T, Solachidis V, Vretos N, *et al*. Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets [J]. Information, 2024, 15(3): 135.

[12] Wang Kai, Peng Xiaojiang, Yang Jianfei, *et al*. Region attention networks for pose and occlusion robust facial expression recognition [J]. IEEE Trans on Image Processing, 2020, 29: 4057-4069.

[13] Huang Qionghao, Huang Changqin, Wang Xizhe, *et al*. Facial expression recognition with grid-wise attention and visual transformer [J]. Information Sciences, 2021, 580: 35-54.

[14] Ghadai C, Patra D, Okade M. A novel facial expression recognition model based on harnessing complementary features in multi-scale network with attention fusion [J]. Image and Vision Computing, 2024, 149: 105183.

[15] Liu Chang, Hirota K, Dai Yaping. Patch attention convolutional vision transformer for facial expression recognition with occlusion [J]. Information Sciences, 2023, 619: 781-794.

[16] Zhao Zengqun, Liu Qingshan, Wang Shanmin. Learning deep global multi-scale and local attention features for facial expression recognition in the wild [J]. IEEE Trans on Image Processing, 2021, 30: 6544-6556.

[17] Gao Shanghua, Cheng Mingming, Zhao Kai, *et al*. Res2Net: a new multi-scale backbone architecture [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662.

[18] Sanghyun W, Jomgchan P, Joon-young L, *et al*. CBAM: convolutional block attention module [C]// Proc of European Conference on Computer Vision. Berlin: Springer, 2018: 3-19.

[19] Liu Hanwei, Cai Huiling, Lin Qingcheng, *et al*. Adaptive multilayer perceptual attention network for facial expression recognition [J]. IEEE Trans on Circuits and Systems for Video Technology, 2022, 32(9): 6253-6266.

[20] Gera D, Balasubramanian S. Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition [J]. Pattern Recognition Letters, 2021, 145: 58-66.

[21] Xiao Junhao, Gan Chenquan, Zhu Qingyi, *et al*. CFNet: facial expression recognition via constraint fusion under multi-task joint learning network [J]. Applied Soft Computing, 2023, 141: article ID 110312.

[22] He Zheng, Meng Bin, Wang Lining, *et al*. Global and local fusion ensemble network for facial expression recognition [J]. Multimedia Tools and Applications, 2023, 82(4): 5473-5494.

[23] Tan Yumei, Xia Haiying, Song Shuxiang. Learning informative and discriminative semantic features for robust facial expression recognition [J]. Journal of Visual Communication and Image Representation, 2024, 98: article ID 104062.

[24] Xu Wei, Wan Yi. ELA: efficient local attention for deep convolutional neural networks [EB/OL]. (2024-03-02). https://arxiv.org/abs/2403.01123.

[25] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al*. Deep residual learning for image recognition [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 770-778.

[26] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: transformers for image recognition scale [C]// Proc of International Conference on Learning Representations. 2021.

[27] Li Shan, Deng Weihong, Du Junpeng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition [J]. IEEE Trans on Image Processing, 2019, 28(1): 356-370.

[28] Barsoum E, Zhang Cha, Ferrer C C, *et al*. Training deep networks for facial expression recognition with crowd-sourced label distribution [C]// Proc of the 18th ACM International Conference on Multimodal. New York: ACM Press, 2016: 279-283.

[29] Selvaraju R R, Cogswell M, Das A, *et al*. Grad-CAM: visual explanations from deep networks via gradient-based localization [J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[30] Hui Ding, Peng Zhou, Chellappa/R. Occlusion-adaptive deep network for robust facial expression recognition [C]// Proc of IEEE International Joint Conference on Biometrics. Piscataway, NJ: IEEE Press, 2020: 1-9.

[31] Zhang Ziyang, Tian Xiang, Zhang Yuan, *et al*. Enhanced discriminative global-local feature learning with priority for facial expression recognition [J]. Information Sciences, 2023, 630: 370-384.