关键点引导与显著帧增强的情感识别网络

黄 忠[1,2+],张丹妮[1],任福继[3],胡 敏[2],刘 娟[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,合肥 230009
3.电子科技大学 计算机科学与工程学院,成都 610056
+ 通信作者 E-mail:huangzhong3315@163.com

摘 要:针对表情与姿态两类情感线索空间占比差异及时间峰值异步的问题,提出一种关键点引导与显著帧增强的情感识别网络。在空间关键点引导子网中,为捕获帧内表情-姿态的情感相关性和互补性信息,基于跨模态注意力和残差结构构建空间关键点引导机制分别获取表情引导语义和姿态引导语义。在时间显著帧增强子网中,为了减少帧间表情-姿态情感峰值异步带来的冗余信息,根据表情引导语义和姿态引导语义度量情感区分度和离散度,提出时间显著帧增强策略实现引导语义序列的时空特征增强。在FABO和CAER视频数据集上的实验结果表明:提出网络的情感识别准确率分别达到95.31%和89.78%,比基线网络分别提高了11.50和13.66个百分点;与相关方法相比,提出方法在自然场景动态视频数据集和静态图片数据集上均具有较好的情感识别性能。

关键词:面部表情;肢体姿态;情感识别;关键点引导;显著帧增强
文献标志码:A **中图分类号:**TP391.41 doi:10.3778/j.issn.1002-8331.2403-0291

Key-Points Guidance and Significant-Frames Enhancement for Emotion Recognition Network

HUANG Zhong[1,2+], ZHANG Danni[1], REN Fuji[3], HU Min[2], LIU Juan[1]
1.School of Electronic Engineering and Intelligent Manufacturing, Anqing Normal University, Anqing, Anhui 246133, China
2.School of Computer Science and Information, Hefei University of Technology, Hefei 230009, China
3.School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610056, China

Abstract: Aiming at the problems of spatial proportion difference and temporal peak asynchrony of emotion cues between facial expression and bodily posture, a key-points guidance and significant-frames enhancement (KGSE-ER) network is proposed for emotion recognition. In the spatial key-points guidance subnet, to capture the intra-frame emotional correlation and complementary information between facial expression and bodily posture, a spatial key-points guidance mechanism is constructed to obtain facial expression guidance semantics and bodily posture guidance semantics based on a cross-modal attention and a residual structure. In the temporal significant-frames enhancement subnet, to reduce the inter-frame redundant information caused by temporal peak asynchrony between facial expression and bodily posture, the emotional discrimination and dispersion are measured according to the facial expression guidance semantics and the bodily posture guidance semantics, and a temporal significant-frames enhancement strategy is proposed to enhance spatio-temporal features of cue-guided semantic sequences. The experimental results on FABO and CAER video datasets show that the emotion recognition accuracy of the proposed network reaches 95.31% and 89.78% respectively, which is 11.50 and 13.66 percentage points higher than the baseline network. Compared with related methods, the proposed network has better emotion recognition performance on both natural scene dynamic video datasets and static image datasets.

Key words: facial expression; bodily posture; emotion recognition; key-points guidance; significant-frames enhancemen

基金项目:国家自然科学基金(62176084);安徽省自然科学基金(1908085MF195);安徽省高校优秀青年人才支持计划项目(gxyqZD2021122);安徽省教育厅自然科学重点研究项目(2022AH051038,2023AH050490)。

作者简介:黄忠(1981—),男,博士,教授,研究方向为情感计算、自然人机交互;张丹妮(1999—),女,硕士,研究方向为情感计算、计算机视觉;任福继(1959—),男,博士,教授,研究方向为情感计算、自然语言处理;胡敏(1967—),女,博士,教授,研究方向为情感计算、图像处理;刘娟(1984—),女,硕士,副教授,研究方向为表情识别、计算机视觉。

收稿日期:2024-03-19 **修回日期:**2024-07-01 **文章编号:**1002-8331(2025)18-0142-15

随着情感计算理论以及人工智能技术的快速发展，基于视觉的情感识别方法因其易获性、直观性、非接触性、准确性等优点，在精神病理学、公共安全、人机交互等领域发挥了越来越重要的作用[1-2]。近些年，基于面部的表情识别占据了视觉情感识别研究的主体[3-7]。但由于遮挡、角度偏转等客观因素以及人类本能情感抑制的影响，基于面部的情感识别方法在自然场景中的性能有待进一步提升[8-9]。研究表明，基于肢体关键点的姿态信息具有抗干扰性[10]与非欺骗性[11-12]，且与面部表情具有较强的情感相关性与互补性[13-14]。如何利用表情与姿态两类情感线索的互补优势提高自然场景情感识别精度，成为人工情感智能领域新的研究热点[15]。然而，当前表情-姿态情感识别方法一方面由于表情与姿态帧内空间占比差异，同一尺度提取的纹理特征易丢失细粒度的情感语义；另一方面，由于表情与姿态帧间时间峰值异步，同一区间聚焦的峰值片段易包含非峰值的冗余信息。因此，如何克服帧内空间占比差异以及帧间时间峰值异步问题仍是表情-姿态情感识别方法亟待解决的问题。

当前表情-姿态情感识别方法主要分为基于峰值帧的方法和基于视频流的方法。基于峰值帧的方法主要将手工挑选的情感峰值帧输入至深度学习模型并进行情感分类。如EMOTIC[16]将包含表情和姿态的峰值帧图片作为卷积神经网络的输入；ECDNet[17]采用表达激励和抑制机制诱导模型捕捉情感区域；BOA[18]在定位人体区域的基础上估计背景物体的情感贡献；Wu等[19]提取不同尺度和不同线索特征并分析其情感相关性；Qi等[20]将注意力机制嵌入三维卷积网络中以提升情感识别性能。此类整体建模方式能够获取表情-姿态的情感相关性，但两类线索的空间占比差异易导致不同尺度信息的丢失。EMERSK[21]、Soumaya等[22]分别提取表情、姿态特征并将其级联后进行情感分类；Wei等[15]采用自适应平均池化融合表情-姿态时空特征；EmoSec[23]基于前向融合和后向融合的特征融合机制融合不同线索特征。此类表情与姿态独立建模方式虽能够关注各线索的局部语义，但仅通过特征级联实现信息融合忽略了两线索间的情感协同信息。为了挖掘表情与姿态的关联信息，CD-Net[24]利用Tubal Transformer结构捕获两类线索间的交互信息；GRERN[25]基于图卷积神经网络和门控循环单元学习两线索间语义关系；Zhou等[26]提出跨通道和混合特征加权网络融合表情与姿态语义；Le等[27]构造全局-局部注意力机制提取更富情感的注意图。整体而言，基于峰值帧的方法利用融合策略能够获取表情-姿态相关性信息，但一方面表情与姿态的空间占比差异易造成提取的空间纹理特征丢失不同尺度信息[28]；另一方面手工标注峰值帧具有较强主观性且易丢失连续性。因此，此类方法难以满足自然场景下人机交互需求[29]。

基于视频流的方法则是将一组视频序列作为模型的输入并通过挖掘空间特征和时序变化实现情感分类。Barros等[30]分别捕获表情和姿态时空特征并直接级联后进行分类；CMEFA[31]基于典型相关分析构建耦合网络以提取表情-姿态互补信息；BDFEI[32]在提取表情和姿态时空特征基础上，借助多变量方差分析计算干扰因素的影响系数，并采用模糊推理方法实现情感分类；Nguyen等[33]、Aqdus等[34]基于双线性池化的融合策略实现表情与姿态特征交互；CMGCN[35]构建跨模态图卷积网络整合表情-姿态相关性特征；陈彩华[36]利用改进的遗传算法完成表情与姿态特征的融合。此类方法致力于捕获表情-姿态情感相关性，但忽略了对峰值区间的关注，导致提取的特征中包含大量冗余的平静帧信息。为了突出峰值帧在情感识别中的作用，Verma等[37]通过度量各帧的绝对差以筛选显著变化区域；Zhang等[38]基于弱监督方法设计跨模态时间擦除网络以寻找关键情感片段。尽管此类关注情感峰值阶段的方法能够捕获更为显著的时序情感特征，但由于表情变化迅速而细微，姿态变化持久而明显，两者时间异步问题导致同一区间聚焦的峰值片段仍包含大量非峰值冗余信息。因此，基于视频流的方法虽然降低了人工标注成本并适用于自然人机交互场景，但其情感时空表征能力和识别精度仍有待进一步提升。

针对表情和姿态两类情感线索空间占比差异及时间峰值异步问题，本文提出一种关键点引导与显著帧增强的情感识别(key-points guidance and significant-frames enhancement for emotion recognition，KGSE-ER)网络，如图1所示。该网络主要包括空间关键点引导子网和时间显著帧增强子网。在空间关键点引导子网中，为克服表情与姿态空间占比差异导致的尺度信息丢失问题，基于跨模态注意力和残差结构构建空间关键点引导(spatial key-points guidance，SKG)机制，利用关键点的定位信息分别引导和补充表情与姿态两类情感线索。在时间显著帧增强子网中，为了减少表情-姿态情感峰值阶段异步带来的冗余信息，根据各帧引导语义的情感分数计算情感区分度和离散度，提出时间显著帧增强(temporal significant-frames enhancement，TSE)策略分别筛选和增强两类线索的情感显著帧。最后，将级联的两线索增强特征序列作为时间Transformer编码器的输入，并将其输出的时空类别向量输入至多层感知机实现情感分类。本文创新点如下：

(1)结合表情与姿态两类情感线索、图像与关键点两种特征模态间的互补优势，提出一种关键点引导与显著帧增强的表情-姿态情感识别网络KGSE-ER。提出的网络从帧级捕获表情与姿态的潜在情感相关性和从视频级筛选增强情感显著帧，并基于时间Transformer自动学习表情和姿态两线索的时空情感互补特征。在

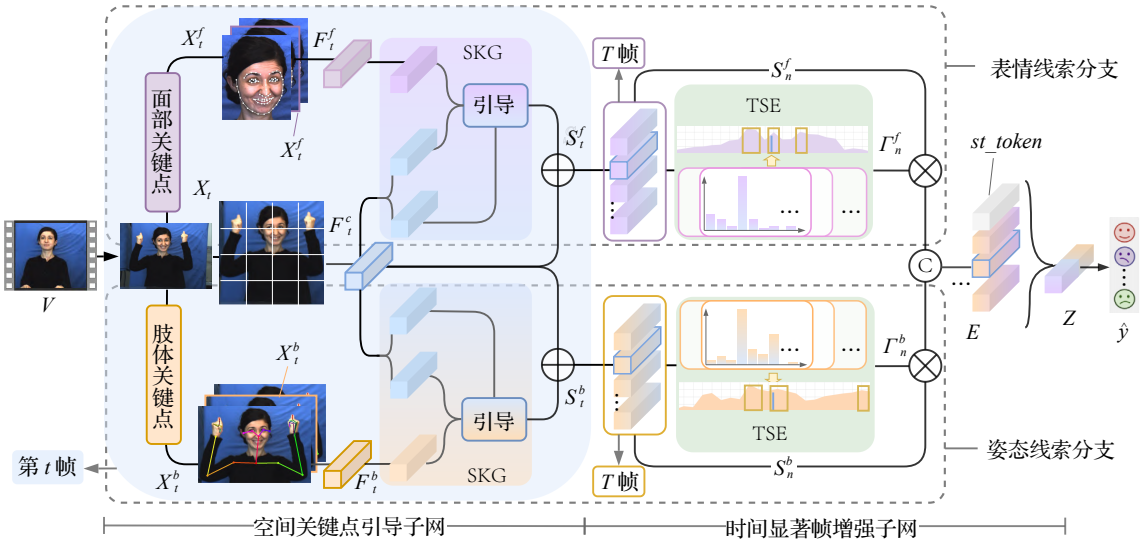


图1　关键点引导与显著帧增强的情感识别(KGSE-ER)网络

Fig.1　Key-points guidance and significant-frames enhancement for emotion recognition(KGSE-ER) network

FABO和CAER视频数据集上的实验表明,该网络通过端-端的方式引导空间依赖信息并增强时序显著帧强度,有效提升了自然场景视频情感识别质量。

(2)为克服表情-姿态空间占比差异,利用抗干扰的关键点几何特征对表情-姿态纹理特征进行引导和补充,并基于注意力和残差结构构建SKG机制。该机制在获取线索间相关性的同时引导纹理特征关注面部局部表情细节和姿态空间远距离依赖,因此克服了空间中模态间的互补性与相关性信息丢失问题,从而提高了帧内空间特征的情感表征能力。

(3)为抑制表情-姿态时序峰值异步,根据各帧引导语义的情感分数计算情感区分度和离散度,提出TSE策略分别聚焦表情与姿态两类线索的情感显著帧。嵌入TSE策略的时间显著帧增强子网通过度量帧级情感显著性与视频级情感关联性,筛选并增强情感显著帧语义信息;同时借助分支结构,克服表情与姿态时序情感异步带来的非峰值帧冗余问题,从而增强了帧间时空特征的情感判别能力。

1　KGSE-ER网络设计

1.1　空间关键点引导子网

鉴于表情与姿态间的情感相关性与互补性,结合两者优势实现双线索情感识别是提升自然场景情感识别质量的重要手段[26]。在视频流中,表情与姿态帧内空间占比差异导致两者尺度大小难以统一,大尺度易丢失局部的表情细节,而小尺度则难以捕获姿态的空间远距离依赖,如图2所示。另外,自然场景下角度偏转、杂乱背景等干扰也增加了纹理特征对情感区域的定位难度[8]。为引导纹理特征关注表情-姿态不同尺度信息,本节在多模态特征提取的基础上构建SKG机制并嵌入空间关键点引导子网。

图2　面部表情与肢体姿态空间占比差异

Fig.2　Difference of spatial proportion between facial expression and bodily posture

1.1.1　多模态特征提取

为获取帧内表情与姿态的情感语义,首先,将输入视频 V 分割成 T 帧图像 $\{\boldsymbol{X}_t\in\Re^{C\times H\times W}\}_{t=1}^{T}$,其中 T、C、H、W 分别表示视频帧数量、通道数、高和宽。然后,分别将各帧图像 $\{\boldsymbol{X}_t\}_{t=1}^{T}$ 馈入ViT[39]网络,获取帧内纹理特征 $\{\boldsymbol{F}_t^c\}_{t=1}^{T}$:

$$\boldsymbol{F}_t^c=f_{\text{image}}(\boldsymbol{X}_t),\ t\in[1,T] \tag{1}$$

式(1)中,$f_{\text{image}}(\cdot)$ 表示ViT网络;$\boldsymbol{F}_t^c\in\Re^{1\times D}$ 为第 t 帧帧内纹理特征,其中 D 为图像嵌入深度。

同时,为了获取抗干扰的关键点定位信息,采用Openpose分别提取各帧面部关键点向量 $\{\boldsymbol{X}_t^f\}_{t=1}^{T}$ 和肢体关键点向量 $\{\boldsymbol{X}_t^b\}_{t=1}^{T}$:

$$(\boldsymbol{X}_t^f,\boldsymbol{X}_t^b)=Openpose(\boldsymbol{X}_t),\ t\in[1,T] \tag{2}$$

式(2)中,$Openpose(\cdot)$ 表示关键点提取器,$\boldsymbol{X}_t^f\in\Re^{2\times k^f}$、$\boldsymbol{X}_t^b\in\Re^{2\times k^b}$ 分别表示第 t 帧面部关键点向量和肢体关键点向量,其中 k^f、k^b 分别表示面部和肢体关键点的数量。然后,为了捕获空间上关键点几何依赖关系,采用DSTA网络[40]提取面部几何特征和肢体几何特征:

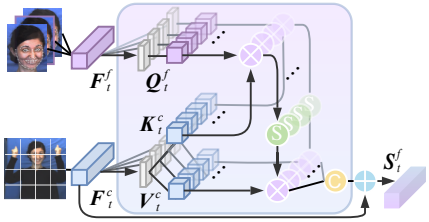
$$\begin{cases}\boldsymbol{F}_t^f=Linear(f_{\text{points}}(\boldsymbol{X}_t^f(t-\mu^f,t+\mu^f))\\ \boldsymbol{F}_t^b=Linear(f_{\text{points}}(\boldsymbol{X}_t^b(t-\mu^b,t+\mu^b))\end{cases},\ t\in[1,T] \tag{3}$$

式(3)中,$Linear(\cdot)$ 表示线性层;$f_{\text{points}}(\cdot)$ 表示DSTA网

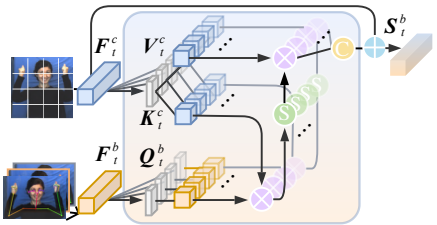
络[40]。μ^f、μ^b分别表示面部和肢体的上下文窗口阈值，当窗口范围超出$[1, T]$时，使用0填充超出窗口对应帧的关键点坐标；$F_t^f \in \Re^{1\times D}$，$F_t^b \in \Re^{1\times D}$分别表示第$t$帧面部几何特征和肢体几何特征。

1.1.2　SKG机制引导

通过多模态特征提取，分别获取了帧内纹理特征$\{F_t^c\}_{t=1}^T$、面部几何特征$\{F_t^f\}_{t=1}^T$和肢体几何特征$\{F_t^b\}_{t=1}^T$。直观地，帧内纹理特征包含表情与姿态全局相关性信息，但易因两者空间占比差异丢失不同尺度信息；同时，面部几何特征和肢体几何特征中包含抗干扰的线索定位信息但缺少局部细节纹理信息。为充分利用几何和纹理特征的互补优势并减少表情-姿态空间占比差异引起的尺度信息丢失，本文采用注意力机制实现关键点几何信息与图像纹理信息的配对，从而引导出模态相关性情感信息；并基于残差结构在保留纹理细节信息的同时关注拓扑依赖关系，从而保留模态互补性信息。构建的SKG机制如图3所示。



(a)面部关键点对帧内纹理特征的引导



(b)肢体关键点对帧内纹理特征的引导

图3　空间关键点引导(SKG)机制

Fig.3　Spatial key-points guidance (SKG) mechanism

首先，以帧内纹理特征$\{F_t^c\}_{t=1}^T$构建多头注意力机制的键向量和值向量，并以面部几何特征$\{F_t^f\}_{t=1}^T$和肢体几何特征$\{F_t^b\}_{t=1}^T$获取多头注意力机制的查询向量：

$$\begin{cases} K_{t,i}^c = F_t^c W_{K,i},\ V_{t,i}^c = F_t^c W_{V,i} \\ Q_{t,i}^f = F_t^f W_{Q,i}^f,\ Q_{t,i}^b = F_t^b W_{Q,i}^b \end{cases},\ t\in[1,T],\ i\in[1,h] \quad (4)$$

式(4)中，$K_{t,i}^c$、$V_{t,i}^c \in \Re^{1\times d_k}$分别为第$t$帧第$i$个注意力头的键向量和值向量，其中$d_k = D/h$，$h$为注意力头数；$Q_{t,i}^f$、$Q_{t,i}^b \in \Re^{1\times d_k}$分别为第$t$帧第$i$个注意力头的表情查询向量和姿态查询向量；$W_{Q,i}^f$、$W_{Q,i}^b$、$W_{K,i}$和$W_{V,i} \in \Re^{D\times d_k}$分别表示可训练的参数矩阵。

然后，采用面部几何特征对帧内纹理特征进行引导并利用残差结构保留全局相关性纹理信息：

$$S_t^f = Concat(head_1, head_2, \cdots, head_h)W_O^f + F_t^c,\ t\in[1,T] \quad (5)$$

其中，$head_i = softmax\left(\frac{(Q_{t,i}^f)(K_{t,i}^c)^{\mathrm{T}}}{\sqrt{d_k}}\right)V_{t,i}^c,\ i\in[1,h]$

式(5)中，$Concat(\cdot)$表示特征级联操作；$softmax(\cdot)$表示归一化指数函数；$W_O^f \in \Re^{hd_k\times D}$为可训练的参数矩阵；$S_t^f \in \Re^{1\times D}$为第$t$帧的表情引导语义。

同理，为了使纹理特征关注姿态拓扑关系，采用肢体几何特征对帧内纹理特征进行引导：

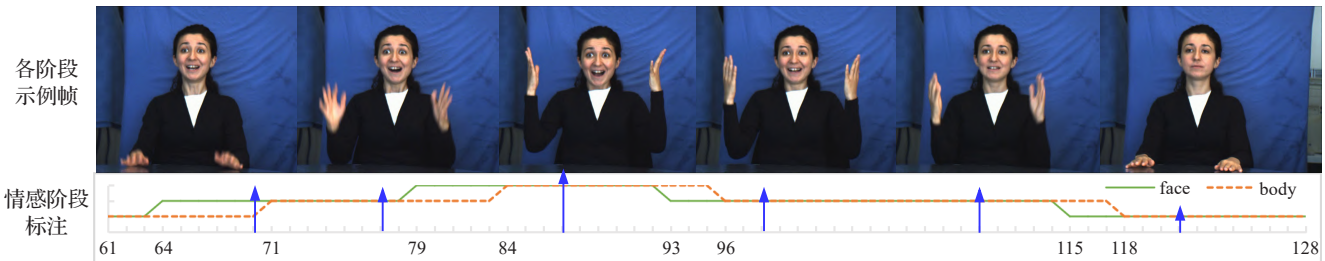
$$S_t^b = Concat(head_1, head_2, \cdots, head_h)W_O^b + F_t^c,\ t\in[1,T] \quad (6)$$

其中，$head_i = softmax\left(\frac{(Q_{t,i}^b)(K_{t,i}^c)^{\mathrm{T}}}{\sqrt{d_k}}\right)V_{t,i}^c,\ i\in[1,h]$

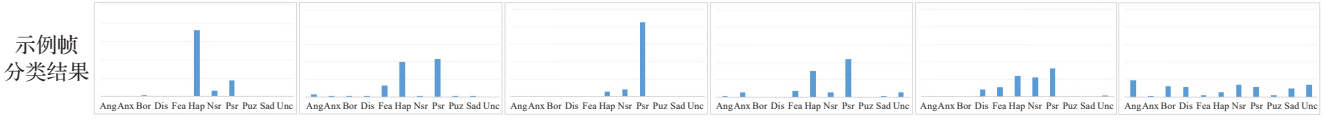
式(6)中，$W_O^b \in \Re^{hd_k\times D}$为可训练的参数矩阵；$S_t^b \in \Re^{1\times D}$为第$t$帧的姿态引导语义。

1.2　时间显著帧增强子网

空间关键点引导子网获取的表情引导语义和姿态引导语义表征了单帧的表情与姿态空间情感信息，但情感变化具有动态性与时序性，如图4所示。图4中，



(a)面部表情和肢体姿态情感阶段的标注(平静-激活-峰值-减弱-平静)



(b)不同阶段示例帧的情感分类结果

图4　“Positive surprise”面部表情和肢体姿态情感阶段比较

Fig.4　Comparison of emotion stages between facial expression and bodily posture for “positive surprise”

“positive surprise”情感包含平静-激活-峰值-减弱-平静等强度变化,且表情与姿态的峰值阶段存在异步。

从图4(a)可以看出,选择大区间情感片段(79~96帧)将保留大量非峰值帧的信息,而选择小区间情感片段(84~93帧)则造成表情或姿态峰值帧的丢失。如何聚焦情感显著帧并抑制情感异步带来的冗余信息是本节需要解决的问题。由图4(b)可知,峰值帧的最高类别分数具有绝对优势,而非峰值帧最高类别分数低且各类别分数差异较小。因此,本文借鉴Yun等[41]采用分类头获取帧级类别激活概率图的思想,设计TSE策略计算各帧的情感显著度,并嵌入时间显著帧增强子网的不同分支以捕获表情-姿态时空情感特征,如图5所示。

1.2.1　TSE策略增强

以表情线索分支显著帧增强为例,为了增强网络对表情峰值帧的聚焦,采用分类头获取各帧表情引导语义 $\{\boldsymbol{S}_t^f\}_{t=1}^T$ 的情感分数 $\{\boldsymbol{score}_t^f\}_{t=1}^T$:

$$\boldsymbol{score}_t^f = softmax(MLP(LN(\boldsymbol{S}_t^f))),\ t\in[1,T] \tag{7}$$

式(7)中,$LN(\cdot)$ 表示层标准化;$MLP(\cdot)$ 表示多层感知机,由两个全连接层和一个GELU激活层组成;$\boldsymbol{score}_t^f = (s_{t,1}^f,\cdots,s_{t,i}^f,\cdots,s_{t,cls}^f)\in\Re^{1\times cls}$ 为第 t 帧表情类别向量,其中 $s_{t,i}^f$ 表示第 t 帧被预测为第 i 类情感的概率分数,cls 为情感类别数。

首先,根据各帧情感分数计算情感区分度 $\{\alpha_t^f\}_{t=1}^T$ 和情感离散度 $\{\beta_t^f\}_{t=1}^T$:

$$\begin{cases}\alpha_t^f = \mathrm{Max}(\boldsymbol{score}_t^f) - \mathrm{Sec}(\boldsymbol{score}_t^f)\\ \beta_t^f = \dfrac{\mathrm{Max}(\boldsymbol{score}_t^f) - \mathrm{Sec}(\boldsymbol{score}_t^f)}{\mathrm{Max}(\boldsymbol{score}_t^f) - \mathrm{Min}(\boldsymbol{score}_t^f)},\ t\in[1,T]\end{cases} \tag{8}$$

式(8)中,Max(·)、Sec(·)和Min(·)分别表示最大值函数、次大值函数和最小值函数;$\alpha_t^f\in\Re^{1\times1}$ 表示第 t 帧的表情区分度,反映了最高分数类别与次高分数类别的相对优势。视频帧的区分度越大,该帧被判别为最高分数类别的置信度越高;反之亦然。$\beta_t^f\in\Re^{1\times1}$ 表示第 t 帧的表情离散度,量化了其余类别间的差异程度。视频帧的离散度越大,该帧最高分数类别与其余类别的混淆程度越低。综合考虑表情区分度和离散度,各帧表情显著度 $\{\boldsymbol{\Gamma}_t^f\in\Re^{1\times1}\}_{t=1}^T$ 可表示为:

$$\boldsymbol{\Gamma}_t^f = \alpha_t^f\beta_t^f,\ t\in[1,T] \tag{9}$$

由式(9)可知,当视频帧的区分度及离散度越大,该帧显著度越大,表明其为显著帧的可信度越高;反之亦然。

然后,为了抑制非峰值片段的影响,根据度量的表情显著度在表情线索分支中筛选前 n 个显著帧:

$$\begin{aligned}&\boldsymbol{\omega}^f, \boldsymbol{Pos}^f = Top(\{\boldsymbol{\Gamma}_t^f\}_{t=1}^T, n)\\ &\boldsymbol{R}^f = Select(\{\boldsymbol{S}_t^f\}_{t=1}^T, \boldsymbol{Pos}^f)\end{aligned} \tag{10}$$

式(10)中,$Top(\cdot)$ 和 $Select(\cdot)$ 分别表示最值索引函数和筛选函数;$\boldsymbol{\omega}^f=[\omega_1^f,\cdots,\omega_i^f,\cdots,\omega_n^f]\in\Re^{n\times1}$ 为 n 个最高的表情显著度组成的权重向量,$\boldsymbol{Pos}^f$ 为相应的显著帧位置序列;$\boldsymbol{R}^f=[\boldsymbol{R}_1^f,\cdots,\boldsymbol{R}_i^f,\cdots,\boldsymbol{R}_n^f]\in\Re^{n\times D}$ 表示以表情显著度筛选的 n 帧引导语义序列。

最后,为突出显著帧在时序片段中的作用,以表情显著度为权重增强 n 帧引导语义序列:

$$\boldsymbol{E}_i^f = \left(\frac{\omega_i^f}{\sum_{j=1}^n\omega_j^f}\times100\right)\boldsymbol{R}_i^f,\ i\in[1,n] \tag{11}$$

式(11)中,$\boldsymbol{E}^f=[\boldsymbol{E}_1^f,\cdots,\boldsymbol{E}_i^f,\cdots,\boldsymbol{E}_n^f]\in\Re^{n\times D}$ 为表情增强

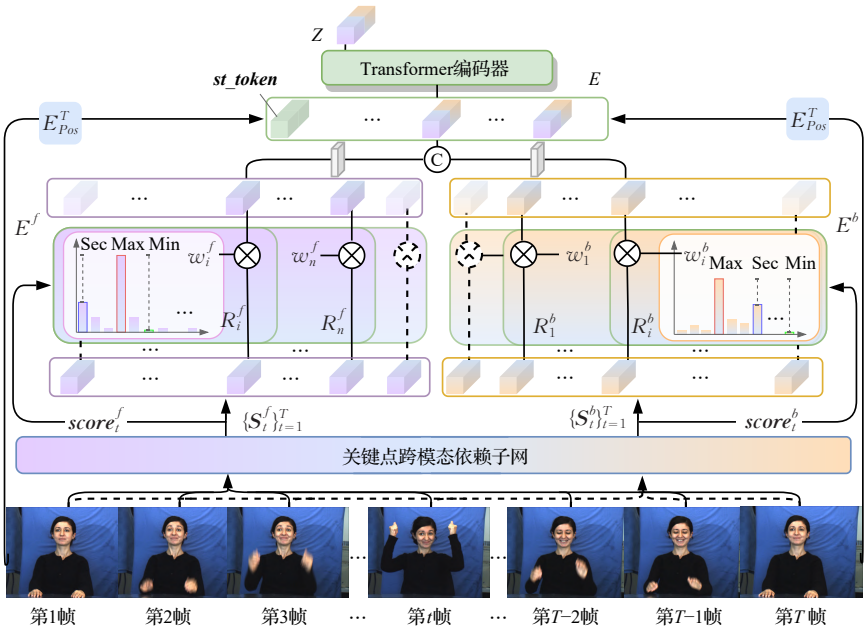


图5　时间显著帧增强(TSE)策略

Fig.5　Temporal significant-frames enhancement(TSE) strategy

特征序列。

按照相同策略，可并行获取表情线索分支的增强特征序列 $E^b=[E_1^b,\cdots,E_i^b,\cdots,E_n^b]\in\Re^{n\times D}$。本文提出的TSE策略通过筛选优质情感显著帧语义序列，剔除了表情、姿态大量非峰值帧冗余信息，并增强了峰值帧的情感显著度；此外，两线索分支独立聚焦各自情感显著帧，克服了两线索在情感变化中的异步问题。

1.2.2　表情-姿态时空特征表示

由TSE策略获取的表情、姿态增强特征序列包含了各自情感线索的峰值情感片段信息。但是，表情情感丰富却易受环境干扰及自身抑制的影响[9]；姿态情感具有非欺骗性但含细粒度情感信息少[11-12]。为捕获表情-姿态间时空互补特征，首先通过线性层将两者维度降至 $D/2$，并级联获得表情-姿态增强特征序列 E：

$$E=Concat(Linear(E^f),Linear(E^b)) \tag{12}$$

式(12)中，$E=[E_1,\cdots,E_i,\cdots,E_n]\in\Re^{n\times D}$。然后，将各帧表情-姿态增强特征 $\{E_i\}_{i=1}^n$ 作为令牌向量，并输入至Transformer编码器获取表情-姿态时空情感特征 $Z\in\Re^{n\times D}$：

$$Z=Encoder((\boldsymbol{st\_token};E)+E_{Pos}^T,l) \tag{13}$$

式(13)中，$Encoder(\cdot,l)$ 表示 l 个串联的Transformer编码器；$E_{Pos}^T\in\Re^{n\times D}$、$\boldsymbol{st\_token}\in\Re^{1\times D}$ 分别表示时间位置编码和时空类别向量。最后，将时空类别向量 $\boldsymbol{st\_token}$ 输入至多层感知机获得最终情感类别向量 $\hat{\boldsymbol{y}}\in\Re^{1\times cls}$：

$$\hat{\boldsymbol{y}}=(\hat{y}_1,\cdots,\hat{y}_i,\cdots,\hat{y}_{cls})=MLP(\boldsymbol{st\_token}) \tag{14}$$

式(14)中，$\hat{y}_i$ 为视频被预测为第 i 类情感的概率；$MLP(\cdot)$ 为多层感知机，与式(7)保持一致。

2　网络优化

为求解KGSE-ER网络参数并加快网络收敛速度，本文采取单分支独立优化和表情-姿态协同的联合优化策略。KGSE-ER网络整体损失可表示为：

$$L=\lambda L^f+(1-\lambda)L^b+L^{st} \tag{15}$$

式(15)中，λ为平衡因子；L^f、L^b分别为表情线索分支、姿态线索分支的独立优化损失函数；L^{st}为表情-姿态协同优化损失函数。各损失函数的计算方式如式(16)所示：

$$\begin{cases}L^f=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}L_{ls}^f(\boldsymbol{y},\boldsymbol{score}_i^f)+L_{ce}^f(\boldsymbol{y},\hat{\boldsymbol{y}}^f)\\ L^b=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}L_{ls}^b(\boldsymbol{y},\boldsymbol{score}_i^b)+L_{ce}^b(\boldsymbol{y},\hat{\boldsymbol{y}}^b)\\ L^{st}=L_{ce}^{st}(\boldsymbol{y},\hat{\boldsymbol{y}})\end{cases} \tag{16}$$

式(16)中，$L_{ls}^f(\cdot)$、$L_{ls}^b(\cdot)$为L-softmax损失函数，从空间上分别度量表情、姿态显著度最高的 n 帧序列的分类结果与真实情感类别间差异；$L_{ce}^f(\cdot)$、$L_{ce}^b(\cdot)$、$L_{ce}^{st}(\cdot)$为交叉熵损失函数，从时间上分别优化表情线索分支、姿态线索分支及表情-姿态协同的情感判别性能；$\hat{\boldsymbol{y}}^f,\hat{\boldsymbol{y}}^b\in\Re^{1\times cls}$分别为表情线索分支和姿态线索分支的情感分类结果，计算方法同式(14)。

3　实验结果与分析

3.1　情感数据集及实验细节

为了说明KGSE-ER网络的识别性能和泛化能力，本文在表情-姿态情感数据集FABO[14]和自然场景情感数据集CAER[42]上进行训练和测试。此外，为了说明SKG机制的有效性，在静态图片数据集FABO[14]、CAER-S[42]和EMOTIC[16]上进行性能分析和相关方法比较。

3.1.1　动态视频情感数据集

FABO[14]是一个表情-姿态“半自发”情感数据集，包含23名被试者554段视频。每段视频为一名被试者循环2~4次的相同情感表达。该数据集标注了11种情感类别，并提供了包括平静-激活-峰值-减弱-平静的情感阶段信息。每个视频被裁剪为2~4个片段(各类别视频片段数如表1所示)，各片段长度从45帧到100帧不等。随机选取18名被试者数据用于训练，其余5名用于测试。

表1　FABO数据集各情感类别视频片段数

Table 1　Number of video clips of each emotion category in FABO dataset

情感类别	视频数	片段数	情感类别	视频数	片段数
Anger	110	268	Ngt surp	22	67
Anxiety	41	97	Pst surp	19	43
Boredom	62	157	Puzzlement	90	286
Disgust	44	113	Sadness	31	67
Fear	43	88	Uncertainty	37	88
Happiness	55	119	总计	554	1 393

注：Ngt surp表示Negative surprise；Pst surp表示Positive surprise。

CAER[42]是一个大规模真实场景情感数据集，包含从79个电视节目中收集的13 201个视频片段，共标注愤怒、厌恶、害怕、开心、难过和惊讶6类情感。视频长度从短片段(约30帧)到长片段(超过120帧)不等。根据文献[42]相同方式，本文采用训练集(70%)、验证集(10%)和测试集(20%)进行实验。

3.1.2　静态图片情感数据集

根据FABO[14]峰值阶段标注，本文采用文献[34]方法提取32 238张峰值帧图片组成FABO静态图片数据集，并随机划分训练集(70%)、验证集(10%)和测试集(20%)。CAER-S[42]为CAER视频数据集的峰值子集，包含从79个电视节目中裁剪的7万幅含有面部和肢体的静态图像，并标注有6种情感以及“中性”类别。本文随机划分训练集(70%)和测试集(30%)进行实验。EMOTIC[16]为常用的真实场景图片情感数据集，共包含从相关数据集选取以及通过谷歌搜索引擎收集的23 571张图像，标注有26种情感类别，并随机划分为训练集(70%)、验证集(10%)和测试集(20%)。

3.1.3 实验细节

为便于KGSE-ER网络的训练以及嵌入模块的性能分析,本文以ViT获取帧内纹理特征、时间Transformer捕获帧间时空特征作为基线网络,并采取式(14)相同多层感知机实现视频情感分类。首先对基线网络进行训练;然后冻结基线网络中ViT的参数,对KGSE-ER网络其他模块参数进行粗训练;最后,基于单分支独立优化与表情-姿态协同优化的联合策略对KGSE-ER网络参数进行微调。此外,为了说明SKG机制的优点,将空间关键点引导子网(窗口阈值 μ^f 和 μ^b 设为0)输出的表情引导语义和姿态引导语义级联,然后输入至式(7)相同多层感知机实现图片情感分类,并在静态图片情感数据集FABO、CAER-S和EMOTIC上与相关方法进行比较分析。KGSE-ER网络输入视频帧大小为224×224,视频帧长度 T 根据不同数据集的视频长度上限分别设为100和120,如截取的帧长不足 T ,将采用0进行填充。实验中,采用Openpose获取70个面部关键点和67个肢体关键点(由25个身体关键点和左右手各21个手部关键点组成)。对于未提取到的关键点坐标,同样采用0进行补齐。本文在Ubuntu18.04操作系统和NVIDIA Tesla V100显卡上对KGSE-ER网络进行训练和测试,主要网络参数设置如表2所示。

表2　KGSE-ER网络参数设置

Table 2　Parameter setting of KGSE-ER network

参数	参考值	参数	参考值
视频帧大小 H, W	224	情感显著帧数量 n	16
视频帧长度 T	100/120	注意力头数 h	4
嵌入深度 D	768	时间编码器数 l	4
面部关键点数 k^f	70	Batch大小	8
肢体关键点数 k^b	67	训练次数	100
面部窗口阈值 μ^f	2	优化器	SGD
肢体窗口阈值 μ^b	3	学习率	0.001

3.2 KGSE-ER网络情感识别效果评价

3.2.1 性能评价

在FABO和CAER数据集上,本文采用情感类别F1值、平均F1值以及准确率Accuracy等指标进行情感识别性能评价,结果如表3所示。

表3中,与基线网络相比,KGSE-ER网络在两个数据集上识别性能均显著提升。在FABO数据集上,KGSE-ER网络的平均F1值和准确率较基线网络分别提高了11.69和11.50个百分点。具体地,在基线网络识别效果最好的“Anger”和“Pst surprise”类别中,KGSE-ER网络的F1值达到了97.65%和97.08%;在基线网络较难识别的“Anxiety”和“Sadness”类别上,KGSE-ER网络的F1值分别提升了25.72和23.01个百分点。在CAER数据集上,KGSE-ER网络的平均F1值和准确率较基线网络分别提高了13.68和13.66个百分点。其中,基线网络较难识别的“Happiness”和“Sadness”两个类别的F1值分别提升了15.17和21.21个百分点。在两个数据集上的实验结果表明,提出的KGSE-ER网络一方面利用SKG机制引导帧内纹理特征,克服了表情-姿态空间占比差异导致的尺度信息丢失问题;另一方面嵌入TSE策略分别聚焦并增强表情与姿态情感显著帧,抑制了表情-姿态峰值异步带来的冗余信息。此外,两个数据集上各类别的F1值均超过82%,这也说明提出的网络具有较好的类别均衡性。

表3　KGSE-ER网络情感识别性能

Table 3　Emotion recognition performance of KGSE-ER network

单位:%

评价指标	情感类别	FABO		CAER	
		Baseline	Ours	Baseline	Ours
F1	Anger	92.75	97.65	77.23	90.79
	Anxiety	64.91	90.63	—	—
	Boredom	81.98	95.57	—	—
	Disgust	77.53	91.91	82.74	93.60
	Fear	84.36	95.13	80.66	89.78
	Happiness	90.72	95.70	71.16	86.33
	Ngt surp	85.71	96.13	70.60	82.75
	Pst surp	94.36	97.08		
	Puzzlement	90.51	98.40	—	—
	Sadness	65.88	88.89	68.40	89.61
	Uncertainty	84.36	94.63	—	—
	Overall	83.01	94.70	75.13	88.81
Accuracy		83.81	95.31	76.12	89.78

注:“—”表示数据集未标注此类情感类别。

3.2.2 消融实验

为了分析本文构建的模块对KGSE-ER网络情感识别性能的影响,本文在FABO和CAER数据集上分别进行6类消融实验,如表4所示。

表4　不同消融策略下的情感识别性能

Table 4　Emotion recognition performance under different ablation strategies

消融策略	模块			准确率/%	
	SKG机制(face)	SKG机制(body)	TSE策略	FABO	CAER
基线网络	×	×	×	83.81	76.12
策略1	√	×	×	88.65	81.39
策略2	×	√	×	85.54	77.04
策略3	√	√	×	90.29	85.22
策略4	×	×	√	88.12	82.50
KGSE-ER网络	√	√	√	95.31	89.78

表4中,策略1和策略2分别在基线网络中嵌入面部关键点引导机制和肢体关键点引导机制;策略3并行捕获表情和姿态引导语义并将其级联后输入至时间Transformer获取表情-姿态时空情感特征;策略4则利用显著帧增强策略聚焦并增强情感显著帧。与基线网络相比,策略1中嵌入面部关键点引导机制使两个数据集的识别准确率分别提升了4.84、5.27个百分点,策略2中嵌入肢体关键点引导机制使准确率分别提升了1.73、

0.92个百分点。这说明提出的关键点引导机制能够充分利用几何语义与纹理语义间的互补优势，引导帧内纹理特征关注表情局部细节或姿态空间远距离依赖。相较于仅关注表情或姿态的策略1和策略2，策略3采用SKG机制分别引导表情局部细节和姿态空间远距离依赖。其在两个数据集上的识别准确率较基线提升了6.48和9.10个百分点。这说明构建的空间关键点引导子网能够有效利用表情与姿态线索间的互补优势，增强单帧表情-姿态空间情感语义的表征能力。与基线网络相比，策略4采用了时间显著帧增强策略，在两个数据集上的识别准确率分别增加了4.31和6.38个百分点。这表明嵌入的时间显著帧增强策略能够聚焦表情和姿态情感的峰值帧，并通过剔除非显著帧和增强显著帧提高识别精度。与相关策略相比，KGSE-ER网络通过嵌入空间关键点引导机制和时间显著帧增强策略，在FABO和CAER数据集上分别达到了95.31%和89.78%的识别准确率，较基线网络提高了11.50和13.66个百分点。这说明提出的KGSE-ER网络能够克服表情和姿态空间占比差异导致的尺度信息丢失以及时间峰值异步造成的非显著帧信息冗余。

3.3　KGSE-ER网络引导效果分析

3.3.1　SKG机制消融策略比较

为说明空间关键点引导(SKG)机制的有效性，本文分别在FABO和CAER数据集上进行了13类消融实验，如表5所示。

表5　SKG机制不同消融策略的比较

Table 5　Comparison of different ablation strategies for SKG mechanism

消融策略	模块	分支		准确率/%	
		face	body	FABO	CAER
策略1	ViT网络	√	√	88.12	82.50
策略2	DSTA网络	√	×	75.54	70.58
策略3		×	√	65.83	61.57
策略4		√	√	78.35	73.76
策略5	ViT网络+DSTA网络(特征加和)	√	×	86.32	79.88
策略6		×	√	82.45	77.13
策略7		√	√	89.71	83.96
策略8	ViT网络+DSTA网络(特征级联)	√	×	89.07	82.84
策略9		×	√	84.74	79.52
策略10		√	√	91.49	85.67
策略11	SKG机制	√	×	92.26	85.45
策略12		×	√	90.17	83.51
策略13		√	√	95.31	89.78

表5中，策略1使用ViT网络获取空间帧内纹理特征；策略2~策略4采用DSTA网络提取空间帧内几何特征；策略5~策略7将ViT网络和DSTA网络的输出特征进行加和操作，而策略8~策略10将两者输出特征级联；策略11~策略13利用关键点几何特征引导图像纹理特征以补充模态互补性与相关性信息。与相关消融策略相比，本文构建的SKG机制在表情线索分支、姿态线索分支以及表情-姿态协同结构上，均具有更好的识别准确率。从模态融合角度，与仅提取图像纹理信息(策略1)或关键点几何关系(策略4)相比，融合两种模态特征的策略7、策略10和策略13具有更好的识别效果。这表明采用图像纹理信息和关键点几何关系两种模态的互补信息实现情感识别的可行性和有效性；与模态特征加和(策略7)、模态特征级联(策略10)相比，本文利用关键点几何特征引导图像纹理特征(策略13)，在两个数据集上的识别准确率分别提升了5.6、5.82和3.82、4.11个百分点。这说明构建的SKG机制通过配对关键点几何信息与图像纹理信息，能够进一步关注到模态间的相关性信息，从而有利于提升情感识别的准确率。从表情线索分支角度，与仅关注表情拓扑关系的DSTA网络(策略2)以及缺失模态相关性信息的特征加和(策略5)和特征级联(策略8)相比，提出的SKG机制(策略11)在两个数据集上的识别准确率分别提高了16.72、14.87个百分点，5.94、5.57个百分点和3.19、2.61个百分点；从姿态线索分支角度，与仅关注姿态远程依赖关系的DSTA网络(策略3)以及缺失模态相关性信息的特征加和(策略6)和特征级联(策略9)相比，提出的SKG机制(策略12)在两个数据集上的识别准确率分别增加了24.34、21.94个百分点，7.72、6.38个百分点和5.43、3.99个百分点。这说明构建的SKG机制能够利用表情或姿态关键点拓扑关系引导图像纹理特征聚焦于关键区域，并捕获帧内模态间的互补性与相关性信息，从而克服了表情-姿态空间占比差异导致尺度信息丢失的问题。因此，构建的SKG机制在表情或姿态单线索情感感知方面仍然具有较好的识别效果。

3.3.2　SKG机制与相关方法比较

为了说明SKG机制的先进性，在FABO的静态图片数据集上与CNN+LSTM[34]、STN[4]、Soumaya等[22]和EMERSK[21]等方法进行比较，如表6所示。

表6　基于FABO图片数据集的相关方法比较

Table 6　Comparison of related methods based on FABO image dataset

方法	提出年份	准确率/%
CNN+LSTM[34]	2021	87.20
STN[4]	2022	84.20
Soumaya等[22]	2022	96.17
EMERSK[21]	2023	96.73
SKG机制	2023	97.33

与仅关注表情线索的STN[4]方法相比，提出的SKG机制充分利用表情与姿态的互补优势，其情感识别准确率提升了13.13个百分点。相较于从裁剪的面部区域和原始图像中提取表情和姿态特征并进行特征级联的CNN+LSTM[34]、Soumaya等[22]和EMERSK[21]等方法，提出的SKG机制利用空间关键点引导帧内纹理特征关注不同尺度信息，其识别准确率分别提升了10.13、1.16和

0.60个百分点。这说明构建的空间关键点引导子网能够克服表情-姿态空间占比差异导致尺度信息丢失的问题,从而提升了表情-姿态情感识别性能。

基于CAER-S图片情感数据集,本文将提出的SKG机制与CAER-Net[42]、GRERN[25]、GLAMOR-Net[27]、MATF[5]、EC-Net[6]、CD-Net[24]、BOA[18]、CAHFW-Net[26]、EMERSK[21]、ECDNet[17]、Wu等[19]方法进行比较,如表7所示。

表7　基于CAER-S图片数据集的相关方法比较

Table 7　Comparison of related methods based on CAER-S image dataset

方法	提出年份	准确率/%
CAER-Net[42]	2019	73.51
GRERN[25]	2021	81.31
GLAMOR-Net[27]	2021	89.88
MATF[5]	2022	86.11
EC-Net[6]	2022	88.01
CD-Net[24]	2022	93.26
BOA[18]	2023	84.82
CAHFW-Net[26]	2023	83.76
EMERSK[21]	2023	92.40
ECDNet[17]	2024	88.06
Wu等[19]	2024	90.83
SKG机制	2024	93.82

在表7中,与CAER-S数据集的基线网络CAER-Net[42]相比,提出的SKG机制识别准确率提升了20.31个百分点。与仅关注表情线索的MATF[5]、EC-Net[6]和ECDNet[17]方法相比,基于表情与姿态双线索的SKG机制准确率分别提高了7.71、5.81和5.76个百分点。与BOA[18]、EMERSK[21]、Wu等[19]方法将各线索独立建模后特征级联相比,提出的SKG机制采用注意力机制和残差结构配对关键点几何信息与图像纹理信息以及保留模态互补性信息,提高了帧内空间特征的情感表征能力。GRERN[25]、GLAMOR-Net[27]、CD-Net[24]、CAHFW-Net[26]等方法虽然构建特征融合模块能够捕获表情与姿态间的相关性,但忽略了自然场景下表情-姿态空间占比差异,因此其情感识别性能均不及SKG机制。总体来说,本文方法在捕获面部-肢体纹理相关性的同时,利用提出的SKG机制关注表情局部细节和姿态空间远距离依赖,克服了因表情-姿态空间占比差异导致尺度信息丢失的问题,因此在自然场景图片数据集CAER-S上获得了最佳的情感识别效果。

此外,在真实场景图片情感数据集EMOTIC[16]上,进一步将提出的SKG机制与EMOTIC[16]、CD-Net[24]、BOA[18]、EmoSec[23]等方法进行比较,如表8所示。由表8中可知,SKG机制的mAP较此数据集上的基线网络EMOTIC[42]提升了10.96个百分点,同时较最先进的方法EmoSec[23]提升了0.37个百分点。这说明提出的SKG机制通过配对关键点几何信息与图像纹理信息,并引导帧内纹理特征关注表情局部细节或姿态空间远距离依赖,有效提升了情感识别的准确率。从各情感类别来看,SKG机制各情感类别的mAP均高于基线的EMOTIC[16]方法;此外,SKG机制在11种情感类别上的mAP优于其他相关方法,尤其在Anticipation、Engagement、Pain、Sympathy等情感类别上,mAP值较当前最优方法提升了33.69、9.15、6.06和11.72个百分点。

表8　基于EMOTIC图片数据集相关方法的mAP值比较

Table 8　Comparison of mAP based on related methods of EMOTIC image dataset

单位:%

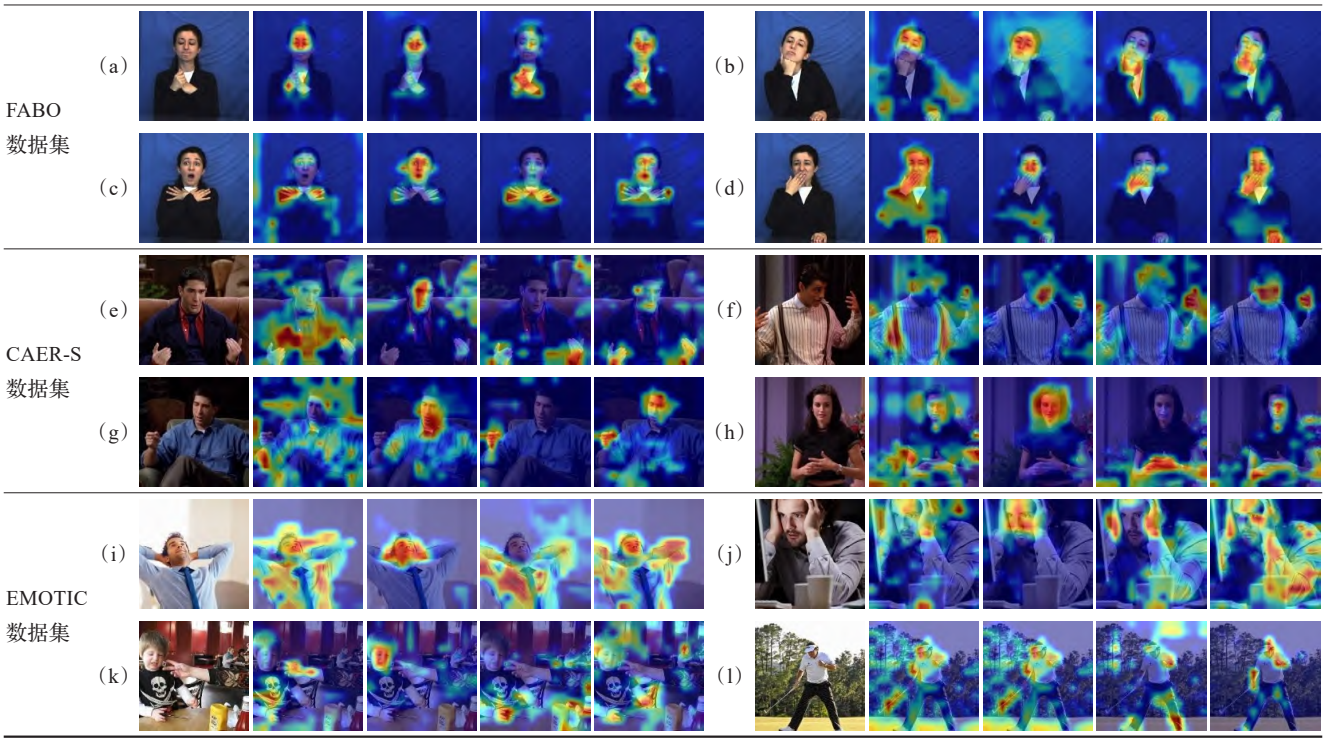
情感类别	EMOTIC[16]	CD-Net[24]	BOA[18]	EmoSec[23]	SKG
Affection	27.85	37.27	37.93	37.81	42.12
Anger	9.49	13.04	13.73	40.58	24.05
Annoyance	14.06	18.92	20.87	27.63	22.21
Anticipation	58.64	56.32	61.08	60.09	94.77
Aversion	7.48	9.82	9.61	88.47	14.94
Confidence	78.35	75.21	80.08	82.32	78.93
Disapproval	14.97	20.11	21.54	63.48	23.17
Disconnection	21.31	30.45	28.32	61.29	38.55
Disquietment	16.89	19.58	22.57	80.20	25.74
Confusion	29.63	21.26	33.50	35.62	37.74
Embarrassment	3.18	2.32	4.16	27.63	15.43
Engagement	87.53	87.01	88.12	41.28	97.27
Esteem	17.73	15.09	20.50	40.51	25.54
Excitement	77.16	68.64	80.11	20.04	82.95
Fatigue	9.70	14.11	17.51	10.70	18.19
Fear	14.14	8.68	15.56	19.61	16.18
Happiness	58.26	77.59	76.01	24.48	81.61
Pain	8.94	11.58	14.56	15.74	21.80
Peace	21.56	26.18	26.76	25.87	31.23
Pleasure	45.46	49.48	55.64	26.32	55.78
Sadness	19.66	39.43	30.80	19.48	32.62
Sensitivity	9.28	11.34	9.59	28.07	17.05
Suffering	18.84	42.35	30.70	29.22	33.09
Surprise	18.81	7.75	17.92	27.38	19.87
Sympathy	14.71	12.28	15.26	21.08	32.80
Yearning	8.34	8.59	10.11	29.63	13.10
mAP	27.38	30.17	32.41	37.87	38.34

3.3.3　SKG机制引导效果可视化

为说明SKG机制的引导效果,本文采用Grad-CAM[43]分别对基线网络、面部关键点引导机制、肢体关键点引导机制以及空间关键点引导子网的关注效果进行热力图可视化,如表9所示。表9中,各热力图激活特征来源如下:“基线网络”为ViT的最终类别令牌向量;“表情引导”为面部关键点引导机制输出的表情引导语义;“姿态引导”表示肢体关键点引导机制输出的姿态引导语义;“空间引导”表示空间关键点引导子网输出的表情引导语义和姿态引导语义的级联表示。由表9中可以看出,在FABO的峰值图片上,空间关键点引导子网利用关键点几何关系引导图像纹理信息,能够有效地辅助基线网络定位和补充缺失的面部-肢体情感区域。在表9(a)和(b)中,基线网络主要关注于眉毛和眼睛等区域,而忽略了对嘴巴及手部的注意;而结合面部关键点引导和肢体关键点引导结果,空间关键点引导子网在表9(a)中增

表9　空间关键点引导(SKG)机制引导效果可视化

Table 9　Visualization of guidance effect of spatial key-points guidance(SKG) mechanism



加了对紧抿的嘴巴、紧握的双手的关注，并以此获取了更多“焦虑”情感信息；在表9(b)中将注意力转移到下弯的嘴角和握拳撑下巴的手部，捕获到更多“无聊”的情感细节。类似的，在表9(c)、(d)中，空间关键点引导子网将注意力集中于双眼、嘴巴以及手部动作等关键情感区域，细化了纹理特征的情感关注范围并补充了互补的几何语义。对于存在光线变化、角度偏转等干扰的真实场景图片数据集CAER-S，空间关键点引导机制不仅关注到面部表情与肢体姿态区域，还抑制了冗余信息的干扰。如表9(e)、(f)和(g)中面部与肢体关键点的引导使ViT分别聚焦于面部区域与手部区域，而忽略不含情感的衣物纹理信息；在表9(h)中，空间关键点引导机制在聚焦到双眼、手指等关键纹理语义的同时，补充了基线网络缺失的双手交握等几何语义，并减少了对背景中左下角植物的关注。特别地，在存在遮挡、杂乱环境等干扰的自然场景数据集EMOTIC上，SKG机制能够捕捉到有效情感区域并弥补单线索情感信息的缺失。如表9(i)和(j)中，SKG机制聚焦于手肘和肩部的特征，并以此推断出“焦虑”和“烦恼”等情感。在表9(k)中，SKG机制辅助空间关键点引导子网避开横在画面中间的第二人手臂，而专注于情感表达人物的表情-姿态情感特征提取。此外，在表9(l)表情缺失的自然场景下，由于Openpose未检测到目标人物的面部关键点，空间关键点引导子网通过定位“握拳打气”的姿态信息判别了“激动”的情感。以上可视化结果表明，提出的SKG机制能够充分利用纹理图像与关键点坐标间的互补优势，引导纹理特征聚焦于表情和姿态不同尺度的情感信息，从而克服表情-姿态空间占比差异导致尺度信息丢失的问题。

3.4　KGSE-ER网络增强效果分析

3.4.1　TSE策略显著帧增强效果

为了说明时间显著帧增强(TSE)策略的显著帧聚焦和增强效果，本文基于FABO数据集的情感阶段标注数据，采用显著帧筛选准确率度量情感显著片段中峰值帧的占比。各情感类别的显著帧增强效果如表10所示。由表10可知，TSE策略在表情与姿态两类线索分支的显著帧筛选准确率分别为81.15%和78.19%，且各情感类别的筛选准确率均高于70%。这说明基于情感显著度的TSE策略能够聚焦表情与姿态两类线索的显著

表10　TSE策略显著帧增强效果

Table 10　Significant-frames filtering effect of TSE strategy

情感类别	显著帧筛选准确率/%	
	表情线索分支	姿态线索分支
Anger	84.42	76.50
Anxiety	78.49	70.13
Boredom	75.74	77.13
Disgust	81.82	75.05
Fear	79.29	72.06
Happiness	91.67	78.76
Ngt surp	78.95	86.59
Pst surp	88.05	79.27
Puzzlement	77.83	90.24
Sadness	82.21	71.80
Uncertainty	74.18	82.61
Overall	81.15	78.19

帧。由两类线索分支的统计结果可知,表情线索的筛选准确率在8种情感类别上均高于姿态线索;但由于面部情感变化不明显而肢体动作幅度大、持续时间长,姿态线索在“Boredom”“Ngt surp”“Puzzlement”“Uncertainty”等情感类别上的显著帧筛选准确率分别高出表情线索1.39、7.64、12.41和8.43个百分点。这表明不同情感类别所呈现的表情线索和姿态线索存在一定差异,且两种线索情感上下文存在互补性。因此,本文在表情线索分支与姿态线索分支中分别嵌入TSE策略聚焦各自的情感峰值帧,从而克服了表情与姿态时序情感异步带来的非峰值帧冗余问题。

3.4.2　TSE策略相关方法比较

为了说明TSE策略的有效性与先进性,本文在FABO和CAER数据集上与平均池化、最大池化、3D-CNN、时间Transformer等典型方法进行比较,如表11所示。

表11　基于TSE策略的相关方法比较

Table 11　Comparison of related methods based on TSE strategy

相关方法	分支		准确率/%	
	face	body	FABO	CAER
平均池化	√	×	82.47	76.72
	×	√	82.93	75.02
	√	√	84.65	77.46
最大池化	√	×	84.13	77.82
	×	√	82.20	74.36
	√	√	85.92	78.53
3D-CNN	√	×	86.23	79.06
	×	√	83.78	75.62
	√	√	87.11	80.27
时间Transformer	√	×	88.65	81.39
	×	√	85.54	77.04
	√	√	90.29	85.22
TSE策略	√	×	92.26	85.45
	×	√	90.17	83.51
	√	√	95.31	89.78

由表11可知,与相关方法相比,提出的TSE策略在表情线索分支、姿态线索分支以及表情-姿态协同结构上均具有较好优势,且双分支的识别准确性高于单线索分支。这说明本文采用的双分支协同结构能够充分利用表情和姿态间情感互补语义提升情感识别质量。特别地,与忽略时序变化特性的平均池化方法以及缺失动态信息的最大池化方法相比,提出的TSE策略筛选并增强情感显著帧,在两个数据集上的识别准确率提升了10.66、12.32个百分点和9.36、11.25个百分点;与仅考虑局部语义的3D-CNN方法相比,提出的TSE策略在剔除情感冗余帧的基础上捕捉情感显著帧的上下文信息,在两个数据集上的识别准确率分别提升了8.20和9.51个百分点。与关注全局时序相关性的时间Transformer方法相比,提出的TSE策略在两个数据集上的识别准确率分别提升了5.02和4.56个百分点。这表明基于情感显著度的TSE策略能够分别聚焦表情与姿态两类线索的情感显著帧,并增强情感显著帧的语义信息,从而抑制了非峰值冗余帧的干扰。

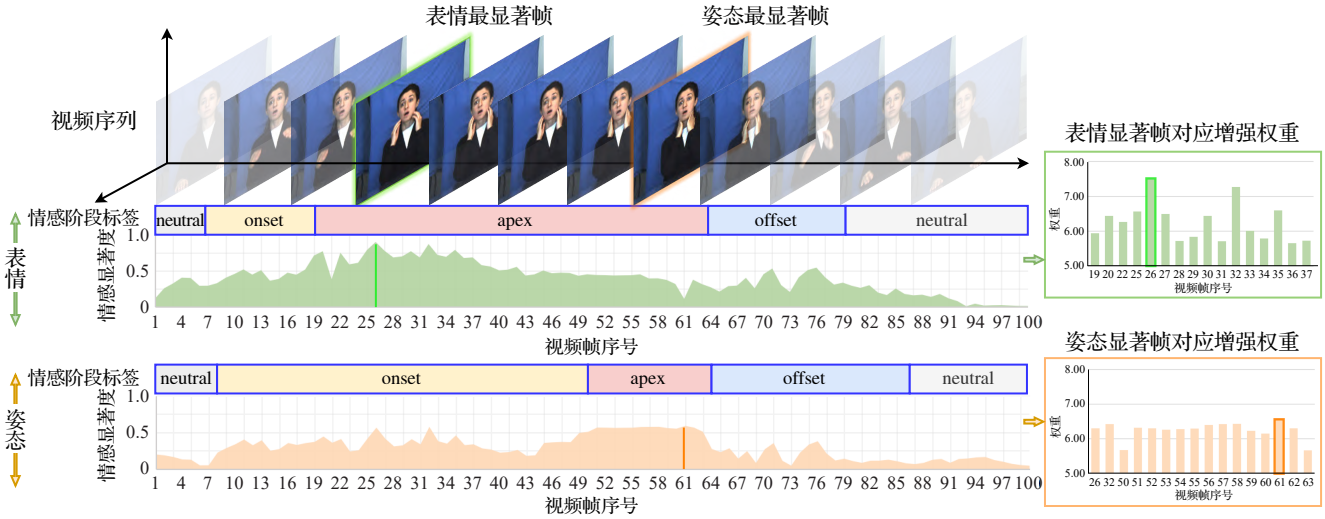
3.4.3　TSE策略增强效果可视化

为了说明时间显著帧增强策略的增强效果,本文将TSE策略计算的各帧表情-姿态显著度与FABO数据集提供的表情-姿态情感阶段标注进行比较,并显示了TSE策略筛选的表情、姿态显著帧及其对应的增强权重,如图6所示。从图6(a)可以看出,面部表情激活(onset)阶段短暂且变化迅速,肢体姿态激活(onset)阶段持续时间长且运动持久;肢体姿态峰值(apex)阶段开始晚且持续时间短,与面部表情峰值(apex)阶段存在明显异步。从表情线索来看,筛选的最高情感显著度(>0.6)片段为第19~37帧,包含在峰值(apex)阶段标签数据(第19~60帧)中,且度量的最高显著帧(第26帧)与人类视觉感知相吻合。从姿态线索来看,筛选的情感显著帧为第50~63帧,与姿态峰值阶段标签数据(第50~64帧)保持一致,筛选的最高显著帧(第61帧)同样符合人类视觉感知效果。与图6(a)样本不同,图6(b)中面部表情激活(onset)阶段较长,肢体姿态激活(onset)阶段较短且手和肘移动迅速;面部表情峰值持续时间较短且变化迅速,而肢体姿态峰值维持时间明显久于面部表情。从表情线索来看,筛选的最高情感显著度(>0.6)片段为第34~48帧,与峰值(apex)阶段标签数据(第33~45帧)基本一致,且度量的最高显著帧(第40帧)与人类视觉感知相吻合。从姿态线索来看,筛选的情感显著帧为第30~47帧,包含在姿态峰值阶段标签数据(第21~49帧)中,筛选的最高显著帧(第36帧)同样与人类视觉感知效果一致。图6中两个样本的情感显著度曲线说明,提出的TSE策略能够聚焦表情和姿态各自的峰值片段,从而抑制了非显著帧的干扰。同时,图6(a)和图6(b)中右侧的柱状图表明:TSE策略实现了表情和姿态的筛选,特别增强了各显著帧的权重。此外,表情显著帧的权重高于姿态显著帧,这也说明表情线索在情感识别中占主导地位而姿态线索起辅助作用。

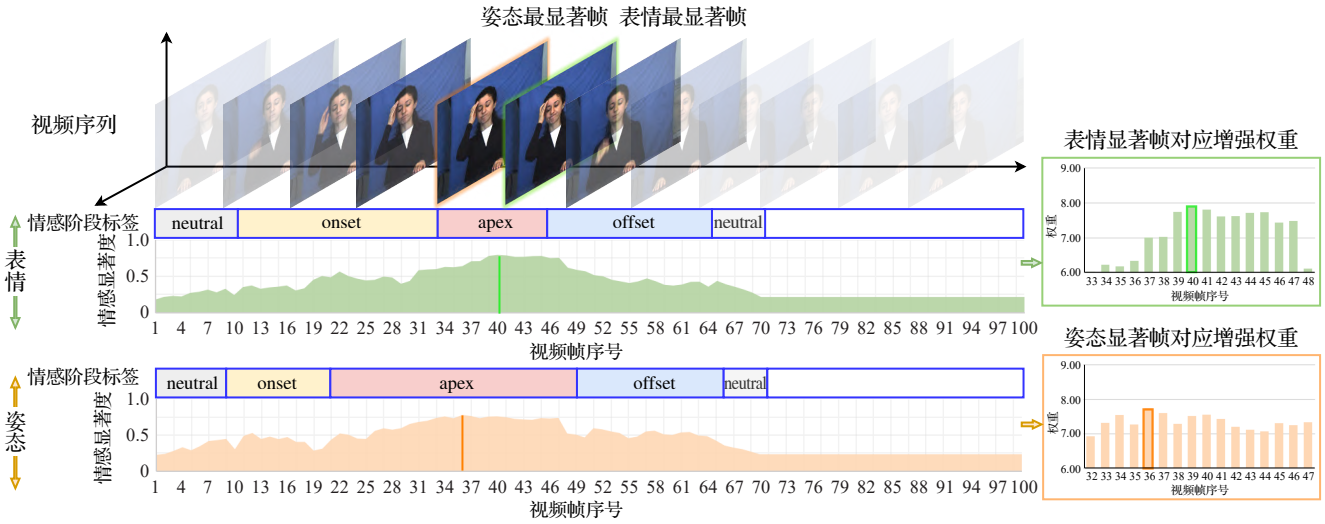
3.5　KGSE-ER网络相关方法比较

在FABO视频数据集上,本文将提出的KGSE-ER网络与CCCNN[30]、C3D+MCB+DBN[33]、CNN+LSTM[34]、Verma等[37]、CMEFA[31]、AM-3D CNN[20]、BDFEI[32]等方法进行比较,如表12所示。

在表12中,与相关方法相比,KGSE-ER网络在FABO视频数据集上取得了最佳的情感识别准确率。与关注表情-姿态整体时空一致性的AM-3D CNN[20]方法、直接级联表情与姿态时空特征的CCCNN[30]方法,以及基于双线性池化融合表情与姿态时空特征的C3D+MCB+DBN[33]、CNN+LSTM[34]方法相比,KGSE-ER网络嵌入空间关键点引导机制克服了空间占比不同导致的



(a)“Ngt surp”情感时间显著帧增强效果可视化



(b)“Puzzlement”情感时间显著帧增强效果可视化

图6　时间显著帧增强(TSE)策略增强效果可视化

Fig.6　Visualization of enhancement effect of temporal significant-frames enhancement(TSE) strategy

表12　基于FABO视频数据集的相关方法比较

Table 12　Comparison of related methods based on FABO video dataset

方法	提出年份	准确率/%
CCCNN[30]	2016	93.65
C3D+MCB+DBN[33]	2018	92.24
CNN+LSTM[34]	2021	94.41
Verma 等[37]	2021	93.40
CMEFA[31]	2023	93.58
AM-3D CNN[20]	2024	88.81
BDFEI[32]	2024	93.09
KGSE-ER网络	2024	95.31

尺度信息丢失的问题，其准确率分别提升了6.5、1.66、3.07和0.90个百分点。与采用多变量方差分析计算干扰因素系数的BDFEI[32]方法相比，KGSE-ER网络采用关键点引导减少空间非情感信息的干扰，并借助TSE策略抑制时序信息的冗余，其识别准确率提升了2.22个百分点。与仅关注表情-姿态整体峰值帧的Verma等[37]方法相比，KGSE-ER网络嵌入的TSE策略分别聚焦并增强了表情和姿态显著帧，其准确率提升了1.91个百分点。这说明提出的TSE策略能够筛选显著情感片段，从而减少了表情-姿态峰值阶段异步带来的冗余信息。与耦合网络CMEFA[31]相比，KGSE-ER网络的情感识别准确率提升了1.73个百分点。这表明本文通过捕获面部-肢体的空间相关性以及度量表情-姿态时序显著性，增强了表情-姿态时空特征的情感表征能力。

为了说明KGSE-ER网络的泛化能力，本文进一步将其与CAER-Net[42]、EMOTION-CNN[16]、GRERN[25]、CMGCN[35]、Zhang等[38]、MSCNN[7]等方法在真实场景情感数据集CAER上进行比较，如表13所示。

表13中，与相关方法相比，KGSE-ER网络在CAER视频数据集上同样取得了最佳的情感识别准确率。与仅关注面部情感的MSCNN[7]相比，KGSE-ER网络借助姿态线索补充面部缺失的情感信息，其识别准确率提升了8.58个百分点。与融合表情全局特征和区域特征的CAER-Net[42]、EMOTION-CNN[16]网络相比，本文构建

表13　基于CAER视频数据集的相关方法比较

Table 13　Comparison of related methods based on CAER video dataset

方法	提出年份	准确率/%
CAER-Net[42]	2019	77.04
EMOTION-CNN[16]	2020	84.52
GRERN[25]	2021	86.73
CMGCN[35]	2022	87.23
Zhang等[38]	2023	80.10
MSCNN[7]	2024	81.20
KGSE-ER网络	2024	89.78

SKG机制捕获空间上表情与姿态间的情感相关性,其情感识别准确率分别提升了12.74和5.26个百分点。与基于图卷积的GRERN[25]和CMGCN[35]网络相比,KGSE-ER网络利用空间关键点引导纹理特征关注表情局部细节和姿态空间远距离依赖,其识别准确率分别提升了3.05和2.55个百分点。而相较于Zhang等[38]构建的基于弱监督的时间擦除网络,KGSE-ER网络嵌入TSE策略筛选显著情感片段并剔除非显著帧的冗余信息,其识别准确率增长了9.68个百分点。这表明KGSE-ER网络通过端-端的方式引导空间依赖信息并增强时序显著帧强度,在自然场景视频情感识别任务中仍具有明显优势。

3.6　超参数对网络性能的影响

本文进一步讨论KGSE-ER网络优化损失函数中的平衡因子 $\lambda\in[0,1.0]$(如式(15)中所示),以及TSE策略筛选的情感显著帧数量 $n\in[4,32]$(如式(10)中所示)对KGSE-ER网络识别性能的影响,实验结果分别如图7、图8所示。

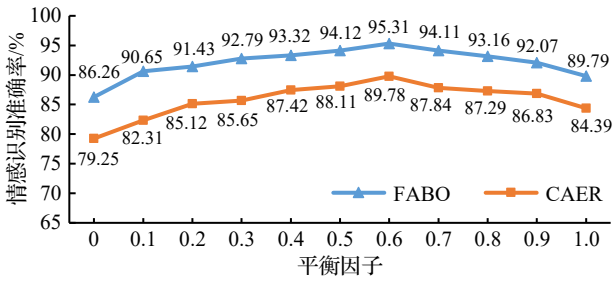


图7　平衡因子对情感识别性能的影响

Fig.7　Effect of balance factor on emotion recognition performance

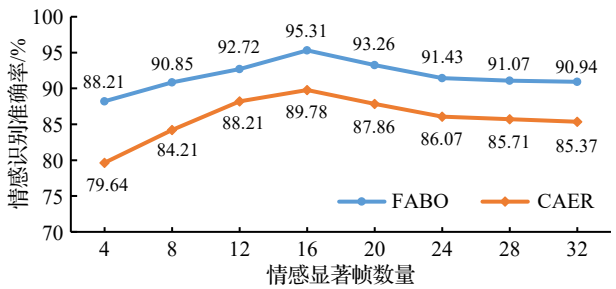


图8　情感显著帧数量对情感识别性能的影响

Fig.8　Effect of number of significant-frames on emotion recognition performance

在图7中,当平衡因子 λ 取0时,KGSE-ER网络对姿态线索分支和表情-姿态协同结构进行联合优化;当 λ 取1.0时,KGSE-ER网络对表情线索分支和表情-姿态协同结构进行联合优化。这两种优化策略由于缺乏表情或姿态的互补信息导致筛选出的显著帧分类错误进而影响识别效果。随着 λ 的增大,式(15)损失函数加大了表情线索分支的重要性,KGSE-ER的情感识别准确率呈上升趋势;当 $\lambda=0.6$ 时,KGSE-ER网络获得表情线索分支和姿态线索分支的最优参数,在两个数据集上均达到了最高的情感识别准确率。在图8中,当 $n<8$ 时,由于筛选的显著帧数量较少,嵌入的TSE策略难以反映情感的动态变化,因此,捕获的表情-姿态时空特征缺乏时序性;当 $n\in[8,16]$ 时,准确率随其增大而呈上升趋势,当 $n=16$ 时,KGSE-ER网络在两个数据集上的识别准确率均达到了峰值;随着 n 的持续增大,情感识别准确率呈下降趋势。主要原因为:保留过多的显著帧虽然能够表征情感的时空动态变化,但包含了大量非峰值帧的冗余信息从而影响了情感识别效果。

3.7　KGSE-ER网络计算性能分析

由于相关方法未提供计算性能方面的数据,本文仅对提出的KGSE-ER网络进行计算性能分析。本文分别统计了SKG机制、TSE策略、基线网络、表情线索分支、姿态线索分支、KGSE-ER网络的参数量、FLOPs以及在两个数据集上的计算开销,如表14所示。

表14　KGSE-ER网络计算性能

Table 14　Computing performance of KGSE-ER network

模块/网络	参数量/10^6	FLOPs/10^8	训练时长/h		测试时长/min	
			FABO	CAER	FABO	CAER
SKG机制	2.56	4.392 5	—	—	—	—
TSE策略	2.86	2.716 3	—	—	—	—
基线网络	43.72	5 427.677 1	7.22	69.10	2.61	26.62
表情线索分支	53.85	5 579.946 5	7.46	71.08	2.78	28.25
姿态线索分支	53.85	5 579.946 5	7.45	71.08	2.78	28.25
KGSE-ER网络	56.70	5 582.943 3	7.51	71.57	2.85	28.88

注:“—”表示该模块未进行单独训练和测试。

由表14可知,本文提出的SKG机制和TSE策略的参数量小于3×10^6、FLOPs小于4.40×10^8。与基线网络相比,KGSE-ER网络的参数量和FLOPs仅有小幅度提升,且由于本文采用了单分支独立优化和表情-姿态协同优化的联合策略,KGSE-ER网络在FABO数据集和CAER数据集上的训练时长较基线网络并无明显提升。此外,在测试性能方面,FABO数据集(含280段测试视频)的测试时长小于3 min,视频段的平均处理时间小于0.62 s,单个视频帧的平均处理时间小于6.2 ms;CAER数据集(含2 640段测试视频)的测试时长小于30 min,视频段的平均处理时间小于0.66 s,单个视频帧的平均处理时间小于6.6 ms。这表明提出的网络能够满足自然场景下实时性能要求。

4　结语

为了提升帧内空间语义和帧间时空特征的情感表征能力，本文提出一种关键点引导与显著帧增强的表情-姿态情感识别网络。在空间关键点引导子网中分别采用面部和肢体关键点几何信息引导帧内纹理特征，构建SKG机制捕获表情局部细节和姿态空间远距离依赖信息，从而克服表情-姿态空间占比差异导致的尺度信息丢失问题；在时间显著帧增强子网中分别度量表情和姿态的情感强度，提出TSE策略筛选两类情感线索的显著帧并抑制非显著帧的冗余信息，从而增强帧间时空特征的判别性。在视频数据集FABO和CAER上分析了提出网络的情感识别效果，通过消融实验和可视化说明了提出的SKG机制和TSE策略的有效性，并讨论了平衡因子 λ 及情感显著帧数量 n 对网络性能的影响。此外，为了说明本文方法的优势，在两个视频数据集上与相关方法进行了比较。实验结果表明，提出的网络在FABO和CAER动态视频数据集上的情感识别准确率分别达到了95.31%和89.78%，较基线网络分别提高了11.50和13.66个百分点；同时，构建的空间关键点引导机制在FABO、CAER-S和EMOTIC静态图片数据集上的识别性能较各自基线网络分别提升了13.33、20.31和10.96个百分点；提出的时间显著帧增强策略在表情与姿态两类线索分支的显著帧筛选准确率分别达到81.15和78.19个百分点。这说明提出网络在自然场景动态视频数据集及静态图片数据集上均具有较好的情感识别效果。此外，网络参数量、FLOPs以及在两个数据集上的计算开销表明，提出网络能够满足自然场景下实时性能要求。然而，自然场景下面部或肢体的部分缺失难以避免，且面部非刚性的瞬时形变和姿态刚性的持续运动存在较大差异。因此，在真实场景中探索这两类情感线索的时空补偿机制将是下一步主要研究工作。

参考文献：

[1] LEONG S C, TANG Y M, LAI C H, et al. Facial expression and body gesture emotion recognition: a systematic review on the use of visual data in affective computing[J]. Computer Science Review, 2023, 48: 100545.

[2] 缪裕青, 董晗, 张万桢, 等. 一种多任务学习的跨模态视频情感分析方法[J]. 计算机工程与应用, 2023, 59(12): 141-147.
MIAO Y Q, DONG H, ZHANG W Z, et al. Cross-modal video emotion analysis method based on multi-task learning[J]. Computer Engineering and Applications, 2023, 59(12): 141-147.

[3] ZHAO R, LIU T S, HUANG Z X, et al. Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition[J]. IEEE Transactions on Affective Computing, 2023, 14(4): 2751-2767.

[4] BARROS P, SCIUTTI A. Across the universe: biasing facial representations toward non-universal emotions with the face-STN[J]. IEEE Access, 2022, 10: 103932-103947.

[5] GUO Y F, HUANG J, XIONG M F, et al. Facial expressions recognition with multi-region divided attention networks for smart education cloud applications[J]. Neurocomputing, 2022, 493: 119-128.

[6] WANG S M, SHUAI H, LIU C G, et al. Bias-based soft label learning for facial expression recognition[J]. IEEE Transactions on Affective Computing, 2023, 14(4): 3257-3268.

[7] 张华忠, 潘曰凯, 涂晓光, 等. 融合三维人脸动态信息和光流信息的人脸表情识别[J]. 计算机科学, 2024, 51(S1): 594-600.
ZHANG H Z, PAN Y K, TU X G, et al. Facial expression recognition integrating 3D facial dynamic information and optical flow information[J]. Computer Science, 2024, 51(S1): 594-600.

[8] 姚鸿勋, 邓伟洪, 刘洪海, 等. 情感计算与理解研究发展概述[J]. 中国图象图形学报, 2022, 27(6): 2008-2035.
YAO H X, DENG W H, LIU H H, et al. An overview of research development of affective computing and understanding[J]. Journal of Image and Graphics, 2022, 27(6): 2008-2035.

[9] SUN N, SONG Y, LIU J X, et al. Appearance and geometry transformer for facial expression recognition in the wild[J]. Computers and Electrical Engineering, 2023, 107: 108583.

[10] LENZONI S, BOZZONI V, BURGIO F, et al. Recognition of emotions conveyed by facial expression and body postures in myotonic dystrophy (DM)[J]. Cortex, 2020, 127: 58-66.

[11] BLYTHE E, GARRIDO L, LONGO M R. Emotion is perceived accurately from isolated body parts, especially hands [J]. Cognition, 2023, 230: 105260.

[12] MAHFOUDI M A, MEYER A, GAUDIN T, et al. Emotion expression in human body posture and movement: a survey on intelligible motion factors, quantification and validation [J]. IEEE Transactions on Affective Computing, 2023, 14(4): 2697-2721.

[13] NOROOZI F, CORNEANU C A, KAMIŃSKA D, et al. Survey on emotional body gesture recognition[J]. IEEE Transactions on Affective Computing, 2021, 12(2): 505-523.

[14] GUNES H, PICCARDI M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior[C]//Proceedings of the 18th International Conference on Pattern Recognition. Piscataway: IEEE, 2006: 1148-1153.

[15] WEI J, HU G Y, YANG X Y, et al. Learning facial expression and body gesture visual information for video emotion recognition[J]. Expert Systems with Applications, 2024, 237: 121419.

[16] KOSTI R, ALVAREZ J M, RECASENS A, et al. Context based emotion recognition using EMOTIC dataset[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(11): 2755-2766.

[17] WANG S M, SHUAI H, ZHU L, et al. Expression complementary disentanglement network for facial expression recognition[J]. Chinese Journal of Electronics, 2024, 33(3): 742-752.

[18] LI W X, DONG X, WANG Y H. Human emotion recognition with relational region-level analysis[J]. IEEE Transactions on Affective Computing, 2023, 14(1): 650-663.
[19] WU S C, ZHOU L, HU Z X, et al. Hierarchical context-based emotion recognition with scene graphs[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 3725-3739.
[20] QI H X, HAN Z F. Emotion recognition and management in the tourism industry during emergency events using improved convolutional neural network[J]. IEEE Access, 2024, 12: 32660-32667.
[21] PALASH M, BHARGAVA B. EMERSK-explainable multimodal emotion recognition with situational knowledge[J]. IEEE Transactions on Multimedia, 2023, 26: 2785-2794.
[22] ZAGHBANI S, BOUHLEL M S. Multi-task CNN for multi-cue affects recognition using upper-body gestures and facial expressions[J]. International Journal of Information Technology, 2022, 14(1): 531-538.
[23] THUSEETHAN S, RAJASEGARAR S, YEARWOOD J. EmoSeC: emotion recognition from scene context[J]. Neurocomputing, 2022, 492: 174-187.
[24] WANG Z L, LAO L J, ZHANG X Y, et al. Context-dependent emotion recognition[J]. Journal of Visual Communication and Image Representation, 2022, 89: 103679.
[25] GAO Q Q, ZENG H X, LI G, et al. Graph reasoning-based emotion recognition network[J]. IEEE Access, 2021, 9: 6488-6497.
[26] ZHOU S W, WU X M, JIANG F, et al. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks[J]. International Journal of Environmental Research and Public Health, 2023, 20(2): 1400.
[27] LE N, NGUYEN K, NGUYEN A, et al. Global-local attention for emotion recognition[J]. Neural Computing and Applications, 2022, 34(24): 21625-21639.
[28] 郭靖圆, 董乙杉, 刘晓文, 等. 注意力机制与Involution算子改进的人脸表情识别[J]. 计算机工程与应用, 2023, 59(23): 95-103.
GUO J Y, DONG Y S, LIU X W, et al. Facial expression recognition based on attention mechanism and involution [J]. Computer Engineering and Applications, 2023, 59(23): 95-103.
[29] 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 M, HU P Y, GE P, et al. Video expression recognition method based on facial motion unit and temporal attention[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(1): 108-117.
[30] BARROS P, WERMTER S. Developing crossmodal expression recognition based on a deep neural model[J]. Adaptive Behavior, 2016, 24(5): 373-396.
[31] CHEN L F, LI M, WU M, et al. Coupled multimodal emotional feature analysis based on broad-deep fusion networks in human-robot interaction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(7): 9663-9673.
[32] LI M, CHEN L F, WU M, et al. Broad-deep network-based fuzzy emotional inference model with personal information for intention understanding in human-robot interaction[J]. Annual Reviews in Control, 2024, 57: 100951.
[33] NGUYEN D, NGUYEN K, SRIDHARAN S, et al. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition[J]. Computer Vision and Image Understanding, 2018, 174: 33-42.
[34] ILYAS C, NUNES R, NASROLLAHI K, et al. Deep emotion recognition through upper body movements and facial expression[C]//Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS-Science and Technology Publications, 2021: 669-679.
[35] LIN B J, LIN Y T, LIU C C, et al. Mental status detection for schizophrenia patients via deep visual perception[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(11): 5704-5715.
[36] 陈彩华. 基于语音、表情与姿态的三模态普通话情感识别[J]. 控制工程, 2020, 27(11): 2023-2029.
CHEN C H. Tri-modal mandarin emotion recognition based on speech, facial expression and body gesture[J]. Control Engineering of China, 2020, 27(11): 2023-2029.
[37] VERMA B, CHOUDHARY A. Affective state recognition from hand gestures and facial expressions using Grassmann manifolds[J]. Multimedia Tools and Applications, 2021, 80(9): 14019-14040.
[38] ZHANG Z C, WANG L J, YANG J F. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 18888-18897.
[39] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J]. arXiv:2010.11929v2, 2020.
[40] SHI L, ZHANG Y, CHENG J, et al. Decoupled spatial-temporal attention network for skeleton-based action recognition[J]. arXiv:2007.03263, 2020.
[41] YUN W L, QI M S, WANG C M, et al. Weakly-supervised temporal action localization by inferring salient snippet-feature[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 6908-6916.
[42] LEE J, KIM S Y, KIM S, et al. Context-aware emotion recognition networks[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 10142-10151.
[43] 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.