DOI：10.13973/j.cnki.robot.240255 **CSTR**：32165.14.robot.240255

基于非对称注意与双池化筛选的情感识别网络

黄 忠[1,2]，张远坤[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，安徽 合肥 230009；
3. 电子科技大学计算机科学与工程学院，四川 成都 610056）

摘 要：为充分利用面部表情与身姿语的情感相关性以及抑制非情感关键帧冗余信息的影响，提出一种基于非对称注意与双池化筛选（asymmetric attention and dual pooling screening，AADPS）的情感识别网络。该网络由帧内空间融合子网和帧间时空融合子网组成。在帧内空间融合子网中，为捕获面部表情和身姿语的空间潜在相关性，构建非对称注意机制，分层获取以手势线索引导的身姿语几何语义和以面部线索引导的帧内情感关联语义。在帧间时空融合子网中，分别采用平均池化和最大池化操作生成视频级情感语义并度量其与帧级语义的相似度，同时设计双池化关键帧筛选策略抑制非峰值帧的影响。实验结果表明：AADPS 网络在 FABO 和 CAER 情感数据集上的情感识别准确率分别达到 94.95% 和 88.69%，相比基线网络提升了 12.22% 和 13.49%。与基于面部表情或身姿语的单模态方法及基于两者情感线索融合的多模态方法相比，本文方法在复杂场景下具有更好的情感识别性能。

关键词：情感识别；面部表情；身姿语；非对称注意；关键帧筛选

Emotion Recognition Network with Asymmetric Attention and Dual Pooling Screening

HUANG Zhong[1,2]，ZHANG Yuankun[1]，REN Fuji[3]，HU Min[2]，LIU Juan[1]

(1. *School of Electronic Engineering and Intelligent Manufacturing*, *Anqing Normal University*, *Anqing* 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: To fully utilize the emotional relevance between facial expression and body language, and to suppress the redundant information of non-emotional key frames, an AADPS (asymmetric attention and dual pooling screening) network for emotional recognition is proposed. The proposed network consists of an intra-frame spatial fusion subnet and an inter-frame spatio-temporal fusion subnet. In the intra-frame spatial fusion subnet, an asymmetric attention mechanism is constructed to capture the potential spatial correlation between facial expression and body language, and to hierarchically obtain the geometric semantics of body language guided by gesture cues and the intra-frame emotional correlation semantics guided by facial cues. In the inter-frame spatio-temporal fusion subnet, the average pooling and max pooling operations are used to generate the video-level emotional semantics and measure the similarity between the video-level emotional semantics and frame-level semantics. Meanwhile, a dual pooling key-frame screening strategy is designed to suppress the influence of non-peak frames. The experimental results demonstrate that the AADPS network achieves emotion recognition accuracies of 94.95% and 88.69% on FABO and CAER datasets, respectively, representing improvements of 12.22% and 13.49% over the baseline network. Compared with the single-modal methods based on facial expression or body language and the multi-modal methods based on the fusion of two emotional cues, the proposed method exhibits superior emotion recognition performance in complex scenarios.

Keywords: emotion recognition; facial expression; body language; asymmetric attention; key-frame screening

情感识别在人机交互、智慧养老、心理健康等“情智兼备”场景中具有重要作用[1-2]。其中，基于面部表情或身姿语[3]（涵盖身体姿态和手势的肢体动作）的视觉情感识别方法因其可访问性、可见性、非接触性和精确性，受到研究者广泛关注[4]。尽管面部表情是表达情感最直接的方式，但在复杂场景中情感识别易受面部遮挡、头部偏转等情形的影响，而且人对情感的主观压抑或掩饰也会降低情感识别准确率；与面部表情相比，身姿语具有抗干扰性、直观性，其在反映原始情感状态方面更具优势[5-6]。鉴于面部表情和身姿语两类情感线索的相关性，如何利用两者的互补优势提升情感识别的准确率，成为当前视觉情感识别领域的研究热点[7-8]。然而，非刚性面部表情和刚性身姿语传递的情感信

基金项目：国家自然科学基金（62176084）；安徽省自然科学基金（1908085MF195）；安徽省教育厅自然科学重点研究项目（2022AH051038）.
通信作者：黄忠，huangzhong3315@163.com 收稿/录用/修回：2024-10-01/2025-02-01/2025-02-10

息存在较大差异，且在时序上伴随大量冗余，如何突出不同情感线索的作用以及抑制非情感关键帧的干扰成为当前情感识别方法亟待解决的问题[9-10]。

当前，基于面部表情和身姿语的情感识别方法主要包括级联方式和关联方式 2 种。1) 基于级联方式的情感识别方法主要将由不同模块提取的情感线索特征串联并进行分类。如 Kosti 等[11] 基于卷积神经网络（CNN）分别提取面部和手势特征并将级联后的语义输入至全连接层实现情感分类；Liu 等[12] 基于 Transformer 网络度量面部表情、全局图像信息和场景信息的情感分数，并聚合 3 种模态分数，生成情感表征向量；在提取面部表情和上半身身姿语特征的基础上，Zaghbani[13]、Palash[14] 采用级联方式实现不同特征的融合。此类面部表情与身姿语独立建模的方式虽能够获取各线索内部的细节信息，但忽略了两类线索间的情感相关性。2) 基于关联方式的情感识别方法则倾向于挖掘面部表情与身姿语之间的情感相关性，并充分利用两者互补信息以提升识别精度。如 Zhou 等[15] 采用交叉注意力机制和跨通道注意力机制，有效捕获面部表情与姿态之间的互补信息；Lin 等[16] 构建跨模态图卷积网络聚合面部表情与姿态的相关性特征；Han 等[17] 将面部表情和手势映射到多特征空间中，并采用模糊多模态融合网络学习两类情感线索的相关性；Gu 等[18] 设计多源学习方法捕获手势和面部表情的模态内自相关性和模态间交叉相关性；Wang 等[19] 构建 CD-Net 网络融合面部表情、姿态以及上下文特征之间的交互信息；Gao 等[20] 在图卷积神经网络基础上采用门控循环单元剔除面部表情和身体姿态间的信息冗余。此外，鉴于情感表达在时间维度上的动态性，李博凯等[21] 提出双分支微表情定位网络搜索微表情峰值帧，Verma 等[22] 通过对比分析连续帧之间的差异从而聚焦情感显著片段。整体而言，上述基于关联方式的情感识别方法能够充分利用面部表情、身姿语两类情感线索的互补性。然而，现有方法主要依赖于传统交叉注意力机制来捕获面部表情和身姿语间的相关性，未能充分考虑不同情感线索的独特性和差异性；同时，现行方法的峰值帧筛选策略未能考虑视频级情感语义与帧级情感语义的一致性，因此难以抑制非关键帧冗余信息的干扰。

1　基于非对称注意与双池化筛选的情感识别网络（Emotion recognition network with asymmetric attention and dual pooling screening）

为了捕获面部表情和身姿语的潜在情感相关性以及聚焦情感显著帧，提出一种基于非对称注意与双池化筛选的情感识别网络，如图 1 所示。提出的 AADPS 网络主要由基于非对称注意力机制（AAM）的帧内空间融合子网和嵌入双池化关键帧筛选策略（DPS）的帧间时空融合子网组成。

1.1　基于非对称注意力引导机制的帧内空间融合子网

相关研究表明：面部表情和身姿语在传递情感信息时存在紧密的关联[23-24]。如人们在高兴时“手舞足蹈”、恐惧时“紧缩双肩”、激愤时“振臂高呼”等。一般地，面部表情包含较丰富的情感信息，但其易受光照、头部偏转、遮挡等因素的影响；而与面部表情具有较强情感相关性的身姿语不仅具

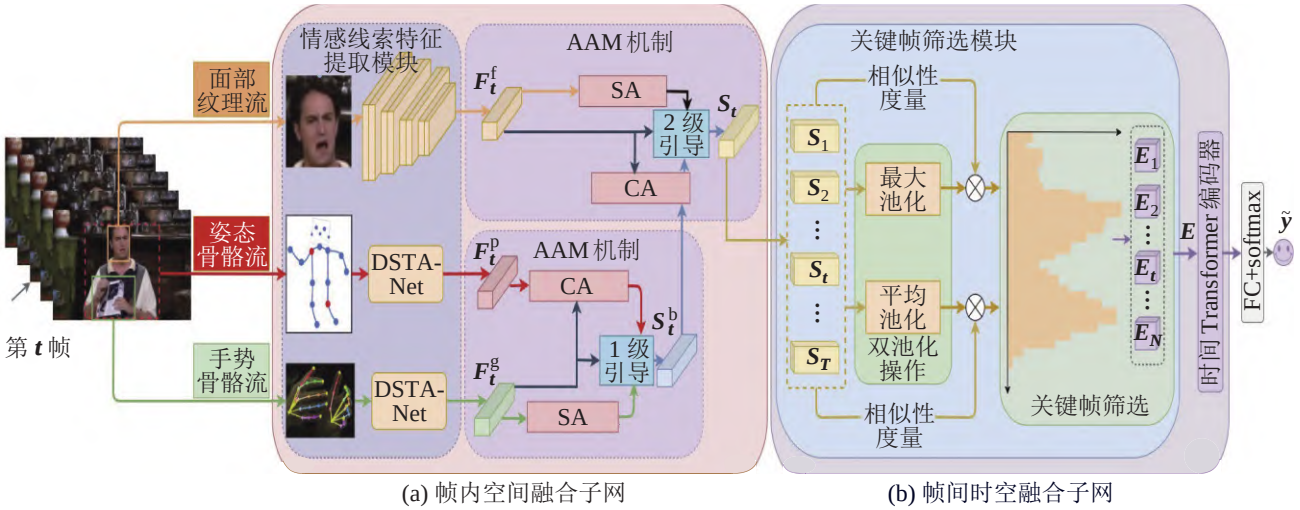


图 1　基于非对称注意与双池化筛选的情感识别网络框架

Fig.1　Emotion recognition network framework based on asymmetric attention and dual pooling screening

备非欺骗性，而且对光照和遮挡保持较高鲁棒性。为了保留不同线索的情感细节特征和捕获各线索之间的情感相关性，提出一种基于注意力机制和残差结构的非对称注意力机制，并以此构建两级引导结构的帧内空间融合子网。该子网主要由情感线索特征提取模块和基于非对称注意力机制的两级引导结构构成。如图 1(a) 所示。

1.1.1 情感线索特征提取

为了在输入视频 V 中提取不同情感线索的特征，从视频 V 中均匀采集 T 帧图像：

$$[\boldsymbol{X}_t]_{t=1}^{T}=\boldsymbol{\Omega}(V,T) \tag{1}$$

式中，$\boldsymbol{\Omega}(\cdot)$ 表示分割运算符，$\boldsymbol{X}_t$ 表示视频 V 的第 t 帧图像。为简化后文表示，仅以 $\boldsymbol{X}_t$ 为例说明面部、姿态、手势 3 类情感线索的特征提取过程。

为捕获第 t 帧图像的面部特征，首先采用人脸识别器定位 $\boldsymbol{X}_t$ 中的人脸区域：

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

式中，$\boldsymbol{I}(\cdot)$ 表示人脸识别器；$\boldsymbol{X}_t^{\mathrm{f}}\in\mathbb{R}^{W\times H}$ 表示第 t 帧的面部区域；W 和 H 分别表示裁剪区域的宽度和长度。然后，采用 ResNet-50 网络 [25] 提取面部区域的特征向量 $[\boldsymbol{F}_t^{\mathrm{f}}]_{t=1}^{T}$：

$$\boldsymbol{F}_t^{\mathrm{f}}=\boldsymbol{\Phi}(\boldsymbol{X}_t^{\mathrm{f}}),\quad t\in[1,T] \tag{3}$$

式中，$\boldsymbol{\Phi}(\cdot)$ 表示 ResNet-50 网络；$\boldsymbol{F}_t^{\mathrm{f}}\in\mathbb{R}^{1\times D}$ 为第 t 帧面部表情特征，其中 D 为图像嵌入深度。

同时，为了在姿态骨骼流和手势骨骼流中获取骨骼信息以及远距离依赖特征，采用 OpenPose API [26] 分别提取姿态关键点信息 $[\boldsymbol{X}_t^{\mathrm{p}}]_{t=1}^{T}$ 和手势关键点信息 $[\boldsymbol{X}_t^{\mathrm{g}}]_{t=1}^{T}$：

$$(\boldsymbol{X}_t^{\mathrm{p}},\boldsymbol{X}_t^{\mathrm{g}})=\boldsymbol{O}(\boldsymbol{X}_t),\quad t\in[1,T] \tag{4}$$

式中，$\boldsymbol{O}(\cdot)$ 表示关键点信息提取器；$\boldsymbol{X}_t^{\mathrm{p}}\in\mathbb{R}^{J_1\times 2}$ 和 $\boldsymbol{X}_t^{\mathrm{g}}\in\mathbb{R}^{J_2\times 2}$ 分别表示第 t 帧姿态关键点几何坐标和手势关键点几何坐标，其中 J_1 和 J_2 分别表示姿态和手势关键点的数量。然后，采用 DSTA-Net 网络 [27] 分别获取姿态和手势骨骼点特征向量 $[\boldsymbol{F}_t^{\mathrm{p}}]_{t=1}^{T}$、$[\boldsymbol{F}_t^{\mathrm{g}}]_{t=1}^{T}$：

$$\begin{aligned}\boldsymbol{F}_t^{\mathrm{p}}&=\boldsymbol{L}(\boldsymbol{\Theta}(\boldsymbol{X}_t^{\mathrm{p}}(t-\varepsilon^{\mathrm{p}},t+\varepsilon^{\mathrm{p}})))\\ \boldsymbol{F}_t^{\mathrm{g}}&=\boldsymbol{L}(\boldsymbol{\Theta}(\boldsymbol{X}_t^{\mathrm{g}}(t-\gamma^{\mathrm{g}},t+\gamma^{\mathrm{g}})))\end{aligned} \tag{5}$$

式中，$\boldsymbol{F}_t^{\mathrm{p}}\in\mathbb{R}^{1\times D}$、$\boldsymbol{F}_t^{\mathrm{g}}\in\mathbb{R}^{1\times D}$ 分别表示第 t 帧的姿态特征和手势特征，$\boldsymbol{L}(\cdot)$ 与 $\boldsymbol{\Theta}(\cdot)$ 分别表示线性层和 DSTA-Net 网络，ε^{p} 和 γ^{g} 分别表示跟踪姿态和手势的时间窗口大小。

1.1.2 非对称注意力机制

以上情感线索特征提取模块可并行获取帧内面部表情特征向量 $[\boldsymbol{F}_t^{\mathrm{f}}]_{t=1}^{T}$、姿态骨骼点特征向量 $[\boldsymbol{F}_t^{\mathrm{p}}]_{t=1}^{T}$ 以及手势骨骼点特征向量 $[\boldsymbol{F}_t^{\mathrm{g}}]_{t=1}^{T}$。一种可行策略是将独立建模的 3 类情感线索特征向量级联。然而，特征级联策略未能充分考虑线索间的情感关联，也未能充分利用不同线索间的互补优势。另一种可行策略是基于典型交叉注意力机制实现 3 类情感线索的关联学习。然而，身体姿态运动区域远大于面部表情和手势的感兴趣区域，导致 3 类情感线索在空间尺度上存在较大差异，因此典型交叉注意力机制难以捕获面部表情线索的局部细节以及突出手势在身姿语情感表达中的辅助作用。为了聚焦面部表情和手势在情感表达中的作用，本文采用注意力机制和残差结构构建一种非对称注意力机制（见图 2），并以此设计分层结构获取以手势线索引导的身姿语几何语义（1 级引导）和以面部线索引导的帧内情感关联语义（2 级引导），如图 1(a) 所示。

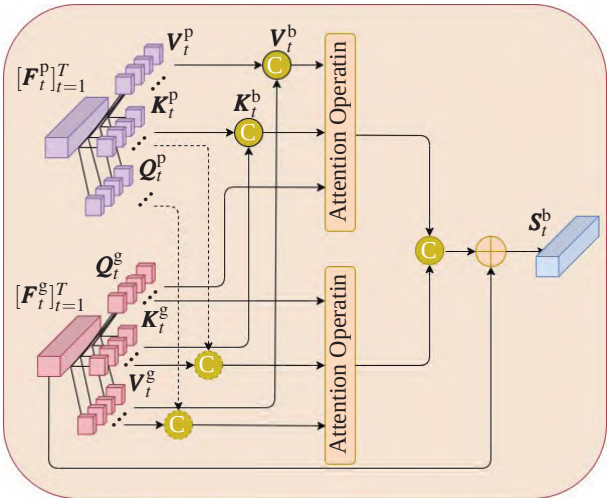


图 2 非对称注意力机制

Fig.2 Asymmetric attention mechanism

在 1 级引导结构中，为突出手势线索在身姿语情感表达中的辅助作用，分别获取姿态骨骼点特征向量 $[\boldsymbol{F}_t^{\mathrm{p}}]_{t=1}^{T}$ 和手势骨骼点特征向量 $[\boldsymbol{F}_t^{\mathrm{g}}]_{t=1}^{T}$ 的查询向量、键向量和值向量，并将二者的键向量和值向量进行拼接得到身姿语的键向量和值向量：

$$\begin{aligned}\boldsymbol{Q}_t^{\mathrm{p}}&=\boldsymbol{F}_t^{\mathrm{p}}\boldsymbol{W}_{\mathrm{Q}}^{\mathrm{p}},\quad \boldsymbol{Q}_t^{\mathrm{g}}=\boldsymbol{F}_t^{\mathrm{g}}\boldsymbol{W}_{\mathrm{Q}}^{\mathrm{g}}\\ \boldsymbol{K}_t^{\mathrm{p}}&=\boldsymbol{F}_t^{\mathrm{p}}\boldsymbol{W}_{\mathrm{K}}^{\mathrm{p}},\quad \boldsymbol{K}_t^{\mathrm{g}}=\boldsymbol{F}_t^{\mathrm{g}}\boldsymbol{W}_{\mathrm{K}}^{\mathrm{g}}\\ \boldsymbol{V}_t^{\mathrm{p}}&=\boldsymbol{F}_t^{\mathrm{p}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{p}},\quad \boldsymbol{V}_t^{\mathrm{g}}=\boldsymbol{F}_t^{\mathrm{g}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{g}}\\ \boldsymbol{K}_t^{\mathrm{b}}&=\boldsymbol{\Gamma}(\boldsymbol{K}_t^{\mathrm{p}},\boldsymbol{K}_t^{\mathrm{g}}),\quad \boldsymbol{V}_t^{\mathrm{b}}=\boldsymbol{\Gamma}(\boldsymbol{V}_t^{\mathrm{p}},\boldsymbol{V}_t^{\mathrm{g}})\end{aligned} \tag{6}$$

式中，$\boldsymbol{Q}_t^{\mathrm{p}}\in\mathbb{R}^{1\times D}$、$\boldsymbol{K}_t^{\mathrm{p}}\in\mathbb{R}^{1\times D}$、$\boldsymbol{V}_t^{\mathrm{p}}\in\mathbb{R}^{1\times D}$ 和 $\boldsymbol{Q}_t^{\mathrm{g}}\in\mathbb{R}^{1\times D}$、$\boldsymbol{K}_t^{\mathrm{g}}\in\mathbb{R}^{1\times D}$、$\boldsymbol{V}_t^{\mathrm{g}}\in\mathbb{R}^{1\times D}$ 分别表示第 t 帧的姿态和手势的查询向量、键向量和值向量，$\boldsymbol{W}_{\mathrm{Q}}^{\mathrm{p}}\in\mathbb{R}^{D\times D}$、

$\boldsymbol{W}_{\mathrm{Q}}^{\mathrm{g}} \in \mathbb{R}^{D\times D}$、$\boldsymbol{W}_{\mathrm{K}}^{\mathrm{p}} \in \mathbb{R}^{D\times D}$、$\boldsymbol{W}_{\mathrm{K}}^{\mathrm{g}} \in \mathbb{R}^{D\times D}$、$\boldsymbol{W}_{\mathrm{V}}^{\mathrm{p}} \in \mathbb{R}^{D\times D}$、$\boldsymbol{W}_{\mathrm{V}}^{\mathrm{g}} \in \mathbb{R}^{D\times D}$ 分别表示可训练的参数矩阵，$\boldsymbol{K}_t^{\mathrm{b}} \in \mathbb{R}^{1\times 2D}$ 和 $\boldsymbol{V}_t^{\mathrm{b}} \in \mathbb{R}^{1\times 2D}$ 分别表示拼接后身姿语的键向量和值向量，$\boldsymbol{\Gamma}(\cdot)$ 表示特征级联操作。

然后，为捕获手势线索和姿态线索两者之间的相关性，采用注意力机制获取手势－姿态交叉注意语义：

$$\boldsymbol{A}_t^{\mathrm{b}} = \boldsymbol{\psi}\left(\frac{\boldsymbol{Q}_t^{\mathrm{g}}\boldsymbol{W}(\boldsymbol{K}_t^{\mathrm{b}})^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}_t^{\mathrm{b}}, \quad t\in[1,T] \tag{7}$$

式中，$\boldsymbol{A}_t^{\mathrm{b}} \in \mathbb{R}^{1\times 2D}$ 表示手势－姿态交叉注意语义；其中，$d_{\mathrm{k}} = D/h$，h 为注意力头数；$\boldsymbol{W} \in \mathbb{R}^{D\times 2D}$ 为可训练的手势查询向量嵌入矩阵，且各帧参数共享；$\boldsymbol{\psi}(\cdot)$ 表示归一化指数函数。

为突出手势线索在身姿语情感表达中的作用，采用自注意力机制捕获手势自注意几何语义：

$$\boldsymbol{A}_t^{\mathrm{g}} = \boldsymbol{\psi}\left(\frac{\boldsymbol{Q}_t^{\mathrm{g}}(\boldsymbol{K}_t^{\mathrm{g}})^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}_t^{\mathrm{g}}, \quad t\in[1,T] \tag{8}$$

式中，$\boldsymbol{A}_t^{\mathrm{g}} \in \mathbb{R}^{1\times D}$ 表示手势自注意几何语义。

最后，将手势自注意几何语义与手势－姿态交叉注意语义进行拼接，并采用残差结构获取身姿语几何语义：

$$\boldsymbol{S}_t^{\mathrm{b}} = \boldsymbol{\Gamma}(\boldsymbol{A}_t^{\mathrm{b}}, \boldsymbol{A}_t^{\mathrm{g}})\boldsymbol{W}_{\mathrm{o}}^{\mathrm{g}} + \boldsymbol{F}_t^{\mathrm{g}}, \quad t\in[1,T] \tag{9}$$

式中，$\boldsymbol{S}_t^{\mathrm{b}} \in \mathbb{R}^{1\times D}$ 表示第 t 帧的身姿语几何语义，$\boldsymbol{W}_{\mathrm{o}}^{\mathrm{g}} \in \mathbb{R}^{3D\times D}$ 为可训练的参数矩阵。由以上计算过程可知，非对称注意力机制提取的身姿语几何语义不仅反映了手势和姿态两种线索的几何相关性，而且突显了手势线索在身姿语情感表达中的辅助作用。

此外，为聚合面部表情和身姿语的相关性语义以及聚焦面部表情线索的关键性作用，采用相似的非对称注意力机制构建 2 级引导结构，如图 1(a) 所示。与手势引导的 1 级结构不同，2 级引导结构以面部表情特征向量为查询向量、身姿语几何语义为键向量和值向量，并采用交叉注意力机制和自注意力机制分别获取面部－身姿语交叉注意语义及面部自注意纹理语义；然后，将面部－身姿语交叉注意语义与面部自注意纹理语义进行拼接，并采用残差结构获取帧内情感关联语义：

$$\boldsymbol{S}_t = \boldsymbol{\Gamma}(\boldsymbol{A}_t^{\mathrm{c}}, \boldsymbol{A}_t^{\mathrm{f}})\boldsymbol{W}_{\mathrm{o}}^{\mathrm{f}} + \boldsymbol{F}_t^{\mathrm{f}}, \quad t\in[1,T] \tag{10}$$

式中，$\boldsymbol{S}_t \in \mathbb{R}^{1\times D}$ 表示第 t 帧的帧内情感关联语义；$\boldsymbol{A}_t^{\mathrm{c}} \in \mathbb{R}^{1\times 2D}$ 和 $\boldsymbol{A}_t^{\mathrm{f}} \in \mathbb{R}^{1\times D}$ 分别表示面部－身姿语交叉注意语义和面部自注意纹理语义，计算方式与式 (7)(8) 中 $\boldsymbol{A}_t^{\mathrm{b}}$ 和 $\boldsymbol{A}_t^{\mathrm{g}}$ 保持一致；$\boldsymbol{W}_{\mathrm{o}}^{\mathrm{f}} \in \mathbb{R}^{3D\times D}$ 为可训练的参数矩阵。

由帧内空间融合子网的构建过程可知，第 1 级结构采用手势线索引导姿态的方式捕获身姿语几何特征，在获取手势与姿态情感相关性的同时充分保留了手势线索的细节语义；第 2 级结构利用面部线索引导身姿语几何特征的方式获取帧内情感关联语义，在捕获面部表情和身姿语互补性信息的同时突出了面部表情线索在情感识别中的显著优势。

1.2　嵌入双池化关键帧筛选策略的帧间时空融合子网

帧内空间融合子网并行获取的帧内情感关联语义 $[\boldsymbol{S}_t]_{t=1}^{T}$ 聚合了面部表情、身姿语的帧内空间语义。然而，人类情感的表达通常呈现出“平静－激活－峰值－减弱－平静”的动态变化过程，其中峰值阶段承载了情感表达的关键信息，而其他时段仅包含较低强度的非关键帧。为突出情感峰值帧的显著作用并减少非峰值帧的干扰，本文提出一种双池化关键帧筛选策略，如图 3 所示。

首先，将帧内情感关联语义 $[\boldsymbol{S}_t]_{t=1}^{T}$ 分别进行平均池化和最大池化操作，并采用 1 维卷积和组归一化生成视频级的平均池化语义和最大池化语义：

$$\begin{aligned}\boldsymbol{M}^{\mathrm{avg}} &= \boldsymbol{\sigma}(\boldsymbol{G}(\boldsymbol{C}(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{S}_t)))\\ \boldsymbol{M}^{\mathrm{max}} &= \boldsymbol{\sigma}(\boldsymbol{G}(\boldsymbol{C}(\max_{1\leqslant t\leqslant T}\boldsymbol{S}_t)))\end{aligned} \tag{11}$$

式中，$\boldsymbol{M}^{\mathrm{avg}}$、$\boldsymbol{M}^{\mathrm{max}}$ 分别表示 T 帧情感语义序列 $[\boldsymbol{S}_t]_{t=1}^{T}$ 的视频级平均池化语义和最大池化语义，$\boldsymbol{C}(\cdot)$ 表示 1 维卷积处理，$\boldsymbol{G}(\cdot)$ 表示组归一化处理，$\boldsymbol{\sigma}(\cdot)$ 表示 sigmoid 函数。

然后，分别度量帧内情感关联语义 $[\boldsymbol{S}_t]_{t=1}^{T}$ 与视频级平均池化语义及最大池化语义间的相似度：

$$\begin{cases}[a_t]_{t=1}^{T} = \boldsymbol{\psi}([\boldsymbol{S}_t]_{t=1}^{T}\otimes\boldsymbol{M}^{\mathrm{avg}})\\ [b_t]_{t=1}^{T} = \boldsymbol{\psi}([\boldsymbol{S}_t]_{t=1}^{T}\otimes\boldsymbol{M}^{\mathrm{max}})\end{cases} \tag{12}$$

式中，$\otimes$ 表示叉乘操作；$[a_t]_{t=1}^{T}$ 和 $[b_t]_{t=1}^{T}$ 分别表示帧级语义与视频级语义间的相似度，a_t 反映了第 t 帧语义与情感整体运动趋势的一致性，b_t 则度量了第 t 帧语义与情感显著片段的差异性。因此，为综合衡量各帧在情感中的作用，以平均池化相似度和最大池化相似度的乘积表示当前帧的情感显著度：

$$\eta_t = a_t b_t, \quad t\in[1,T] \tag{13}$$

由式 (13) 可知，η_t 表示第 t 帧的情感显著度。η_t 越大，表明第 t 帧被判别为关键帧的可信度越高；反之亦然。

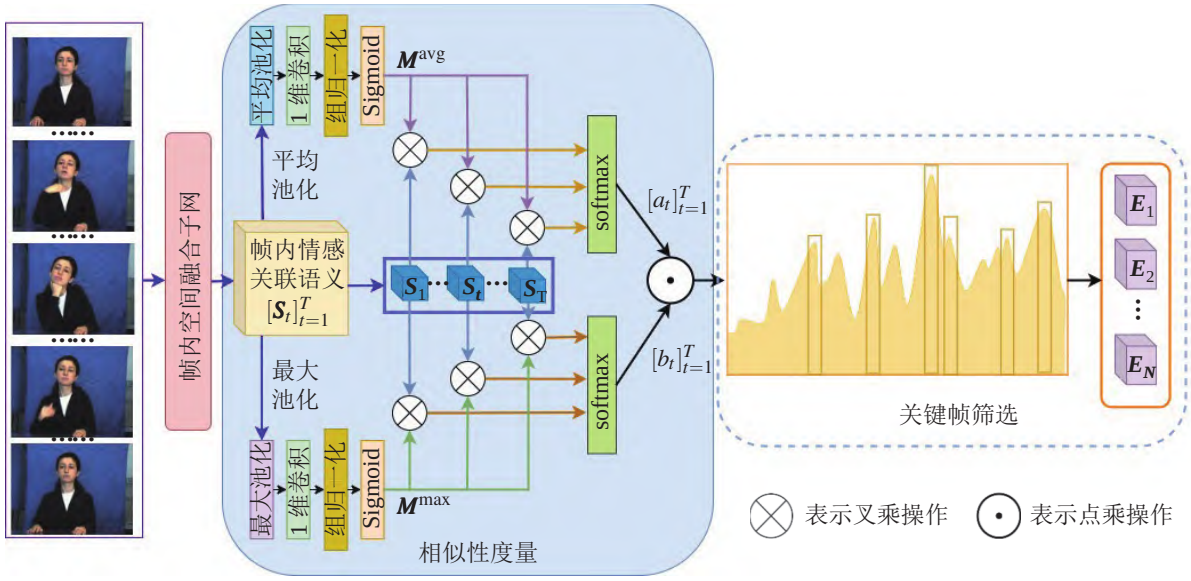


图 3 双池化关键帧筛选策略

Fig.3 Dual pooling key-frame screening strategy

最后，为了抑制非关键帧的影响，根据各帧显著度筛选出前 N 个优质的情感关键帧：

$$\boldsymbol{P}=\boldsymbol{K}([\eta_t]_{t=1}^T,N),\quad \boldsymbol{E}=\boldsymbol{\Upsilon}([\boldsymbol{S}_t]_{t=1}^T,\boldsymbol{P}) \tag{14}$$

式中，$\boldsymbol{K}(\cdot)$、$\boldsymbol{\Upsilon}(\cdot)$ 分别表示最值索引函数和筛选函数，$\boldsymbol{P}=[p_1,\cdots,p_t,\cdots,p_N]\in\mathbb{R}^N$ 表示筛选的 N 帧情感关键帧位置序列，$\boldsymbol{E}=[\boldsymbol{E}_1,\cdots,\boldsymbol{E}_t,\cdots,\boldsymbol{E}_N]\in\mathbb{R}^{N\times D}$ 表示筛选后的帧间情感显著语义序列。

在关键帧筛选模块中，采用双池化关键帧筛选策略尽管保留了优质情感显著帧，但筛选的显著帧间缺乏时序上下文信息。由此，为获取情感关键帧的时空特征，以语义序列 $[\boldsymbol{E}_t]_{t=1}^N$ 为令牌向量并送入 Transformer 编码器：

$$\boldsymbol{Z}=\boldsymbol{T}(\boldsymbol{E},l) \tag{15}$$

式中，$\boldsymbol{Z}\in\mathbb{R}^{N\times D}$ 表示输入视频 V 的帧间时空情感特征，$\boldsymbol{T}(\cdot,l)$ 表示 l 个 Transformer 编码器的级联。

最后，为获取输入视频的最终情感类别，将帧间时空情感特征 $\boldsymbol{Z}$ 输入至多层感知机，最终获得各情感类别的概率分布 $\tilde{\boldsymbol{y}}$：

$$\tilde{\boldsymbol{y}}=\mathrm{softmax}(\boldsymbol{\theta}(\boldsymbol{Z})) \tag{16}$$

式中，$\tilde{\boldsymbol{y}}=(y_1,\cdots y_i,\cdots,y_n)$，$n$ 为可判别的情感类别数量；$\boldsymbol{\theta}(\cdot)$ 表示全连接层。

至此，构建的情感识别网络 AADPS 采用非对称注意力机制及两级引导结构捕获了帧内面部表情和身姿语的潜在相关性，并采用双池化筛选策略实现了情感显著帧的筛选。为求解 AADPS 网络参数，本文以交叉熵损失函数作为目标函数并采用端到端方式完成优化：

$$L=-\sum_{i=1}^{n}\boldsymbol{y}\ln\tilde{\boldsymbol{y}} \tag{17}$$

式中，$\boldsymbol{y}$ 表示真实情感标签，$\tilde{\boldsymbol{y}}$ 表示网络预测的情感类别向量。

2 实验结果及分析（Experimental results and analysis）

2.1 数据集及实验细节

2.1.1 数据集

为了说明 AADPS 网络的识别性能和泛化能力，本文在公开情感数据集 FABO [24] 和 CAER [28] 上进行训练和测试。FABO 是一个包含面部、身体和手势的人类非语言情感行为数据集。其共收集了 23 名受试者近 1 900 个面部表情和身姿语的视频流，标注有 Anger、Ngt surprise、Pst surprise、Fear、Anxiety 和 Boredom 等 11 种情感。CAER 是一个由 13 201 个电视节目视频片段组成的视频情感数据集，并标注有 Anger、Disgust、Fear、Happiness、Sadness、Surprise、Neutral 等 7 种情感状态。在 FABO 数据集上，随机选取 18 名被试者数据用于训练，其余 5 名用于测试。在 CAER 数据集上，采用文 [28] 实验方式，将数据集随机划分为训练集（70%）、验证集（10%）和测试集（20%）。根据以上划分方式，本文在两个数据集上分别进行 10 次随机划分，并将 10 次实验结果的平均值作为性能评价指标。

2.1.2　数据预处理

为确保输入数据符合网络训练要求，在离线阶段采用如下预处理步骤裁剪图像序列，生成身体与骨骼关键点坐标序列。首先将 FABO 和 CAER 数据集中的视频流以 15 帧/s 的帧率转换为图像帧；然后基于 OpenCV 对每一帧图像进行人脸检测和对齐操作，并将人脸区域裁剪为 224×224 大小的图像；最后采用 OpenPose 库[26] 跟踪定位 25 个身体关键点和 42 个手势关键点。本文在 Ubuntu 18.04 操作系统上采用 NVIDIA Tesla V100 显卡对 AADPS 网络进行训练和测试（如表 1 所示）。训练过程中，采用随机梯度下降（SGD）优化器，初始学习率设为 0.000 1。

表 1　AADPS 网络参数设置

Tab.1　Parameter setting of AADPS network

参数	参考值	参数	参考值
视频帧大小 H,W /像素	224	视频长度 T /帧	100
肢体关键点数量	25	关键帧数量 n	12
手势关键点数量	42	注意力头数 h	8
肢体窗口大小 ε^{p} /帧	3	时间编码器数 l	4
手势窗口大小 γ^{g} /帧	3	批量大小	8

2.2　AADPS 网络情感识别性能分析

为了分析 AADPS 网络的情感识别性能，采用情感类别 F1 值（%）、平均 F1 值（%）以及准确率（%）等指标进行评价。在 FABO 和 CAER 数据集上的实验结果如表 2、表 3 所示。

表 2　在 FABO 数据集上 AADPS 网络的情感识别性能

Tab.2　The emotion recognition performance of the AADPS network on FABO dataset

单位：%

情感识别性能		基线网络	AADPS	提升幅度
F1 值	Anger	92.62±0.31	94.93±0.24	02.31
	Anxiety	63.65±0.36	89.75±0.38	26.10
	Boredom	81.88±0.22	92.33±0.27	10.25
	Disgust	78.21±0.67	90.21±0.75	12.00
	Fear	84.36±1.04	94.30±0.84	09.94
	Happiness	90.25±0.59	94.42±0.45	04.17
	Ngt surprise	85.72±0.34	92.85±0.17	07.13
	Pst surprise	93.27±0.24	94.57±0.52	01.30
	Puzzlement	88.92±0.82	96.21±0.60	07.29
	Sadness	66.71±1.13	86.37±0.71	19.66
	Uncertainty	84.38±0.67	92.42±1.01	08.04
	平均 F1 值	82.06±0.64	92.74±0.46	10.68
准确率		82.73±0.51	94.95±0.23	12.22

表 3　在 CAER 数据集上 AADPS 网络的情感识别性能

Tab.3　The emotion recognition performance of the AADPS network on CAER dataset

单位：%

情感识别性能		基线网络	AADPS	提升幅度
F1 值	Anger	76.23±0.67	88.72±0.87	12.49
	Disgust	82.02±0.95	92.35±1.05	10.33
	Fear	78.63±0.52	87.67±0.43	09.04
	Happiness	71.21±1.51	86.13±0.31	14.92
	Surprise	69.27±0.87	80.21±0.27	10.94
	Sadness	67.72±0.89	88.17±0.66	20.45
	平均 F1 值	74.33±1.17	87.32±0.83	12.99
准确率		75.20±0.74	88.69±0.17	13.49

与基线网络 ResNet-50+DSTA-Net 相比，AADPS 网络在两个数据集上识别性能均大幅度提升。在 FABO 数据集上，AADPS 网络的平均 F1 值和准确率较基线网络分别提高了 10.68% 和 12.22%，且各自标准差低于基线网络。具体地，在基线网络识别效果最好的 Anger 和 Pst surprise 类别中，AADPS 网络的 F1 值分别达到了 94.93% 和 94.57%；在基线网络较难识别的 Anxiety 和 Sadness 类别上，AADPS 网络的 F1 值分别提高了 26.10% 和 19.66%。在 CAER 数据集上，AADPS 网络的平均 F1 值和准确率较基线网络分别提高了 12.99% 和 13.49%，且各自标准差小于基线网络。其中，基线网络较难识别的 Surprise 和 Sadness 两个类别的 F1 值分别提升了 10.94% 和 20.45%。

在两个数据集上的实验结果表明，提出的 AADPS 网络一方面利用非对称注意力机制分层获取以手势线索引导的身姿语几何语义和以面部线索引导的帧内情感关联语义，充分发挥了面部表情和身姿语的互补优势；另一方面嵌入双池化关键帧筛选策略获取情感显著帧，有效剔除了非峰值帧携带的冗余信息。此外，在两个数据集上各类别的 F1 值均超过 80%，这也说明提出的网络具有较好的类别均衡性。

2.3　消融实验

2.3.1　非对称注意力机制对网络情感识别性能的影响

为了说明构建的非对称注意力机制对 AADPS 网络情感识别性能的影响，本文分别在 FABO 和 CAER 数据集上设计 5 类消融实验，如表 4 所示（表中的“√”和空白分别表示采用对应模块及未采用对应模块）。在表 4 中，基线网络分别采取面部、面部+姿态、面部+手势、面部+身姿语 4 种输

表 4 不同注意力机制下 AADPS 网络的情感识别性能

Tab.4 Emotion recognition performance of the AADPS network under different attention mechanisms

消融策略	网络输入	模块				准确率 /%	
		CA [29]	CrossViT [30]	AAM（手势－姿态引导）	AAM（面部表情－身姿语引导）	FABO	CAER
基线网络	面部					76.31	67.26
	面部+姿态					79.27	69.54
	面部+手势					80.11	72.97
	面部+身姿语					82.73	75.20
策略 1	面部+身姿语	√				83.71	76.67
策略 2			√			83.86	76.82
策略 3				√		85.17	77.19
策略 4					√	87.34	80.21
策略 5（本文）				√	√	89.28	84.95

入方式，且以 ResNet-50+DSTA-Net 为基线网络实现情感分类；其他策略均以面部+身姿语为输入，其中，策略 1 采用典型交叉注意力机制 [29]（Cross-Attention，CA）实现面部纹理特征对姿态和手势级联特征的引导；策略 2 采用 CrossViT 注意力机制 [30] 实现面部纹理特征对姿态和手势级联特征的引导；策略 3 仅利用 AAM 实现手势线索对姿态引导；策略 4 仅采用 AAM 完成面部纹理特征对姿态和手势级联特征的引导；策略 5 采用本文设计的两级 AAM 获取帧内情感关联语义。在基线网络的 4 种不同输入方式中，面部+身姿语的情感识别准确率最高，且面部+ 手势的识别准确率高于面部+姿态的融合方法。这一方面表明身姿语线索能够有效补充面部表情信息，从而更好地反映情感状态；另一方面说明手势在身姿语情感表达上比姿态更具优势。与以面部+身姿语为输入的基线网络相比，策略 1 和策略 2 分别采用 CA 和 CrossViT 机制实现面部纹理特征对姿态和手势级联特征的引导，在 FABO 和 CAER 数据集上的情感识别准确率分别提高了 0.98%、1.47% 和 1.13%、1.62%；策略 3 采用 AAM 实现手势对姿态的引导，在 FABO 和 CAER 数据集上的情感识别准确率分别提高了 2.44%、1.99%；策略 4 则采用 AAM 实现面部纹理特征对姿态和手势的级联特征进行引导，其在两个数据集上的准确率分别提升了 4.61%、5.01%。这说明提出的 AAM 机制能够在身姿语情感线索中突出手势的引导作用，在多模态线索融合中发挥面部表情的显著优势。策略 1、2、4 同样采用了注意力机制实现面部纹理特征对姿态和手势级联特征的引导，但相较于策略 1、2，策略 4 在两个数据集上的准确率分别提升了 3.63%、3.54% 和 3.48%、3.39%。这表明提出的 AAM 机制较传统交叉注意力机制能够更好地捕获不同情感线索间的相关性。相较于仅采用单一 AAM 的策略 3 和策略 4，策略 5 采用两级 AAM 分层获取以手势线索引导的身姿语几何语义和以面部线索引导的帧内情感关联语义，在两个数据集上情感识别准确率分别提高了 4.11%、7.76% 和 1.94%、4.74%。这说明构建的帧内空间融合子网能够有效利用面部表情和身姿语的互补优势，增强了帧内情感关联语义的表征能力。

2.3.2 双池化关键帧筛选策略对网络情感识别性能的影响

为了说明设计的双池化关键帧筛选策略对 AADPS 网络情感识别性能的影响，本文分别在 FABO 和 CAER 数据集上设计不同关键帧筛选策略的比较实验，如表 5 所示（表 5 中，未引入注意力机制及未涉及相关筛选策略均用空白表示）。与基线网络（ResNet-50+DSTA-Net）类似，策略 1 采用 3D-CNN [28]+DSTA-Net 直接捕获整个视频流特征，并将其输入至分类器进行分类。策略 2～5、策略 6～9 和策略 10～12 分别表示在帧内空间融合子网中嵌入 CA、CrossViT 和非对称注意力机制，其中，策略 2、6、10 不采用关键帧筛选策略；策略 3、7、11 表示仅采用平均池化操作筛选情感显著帧；策略 4、8、12 则表示仅采用最大池化法筛选情感显著帧；策略 5、9、13 表示采用本文提出的 DPS 策略筛选情感显著帧。与基线网络相比，策略 1 对整个视频流不进行额外分帧，其在 FABO 和 CAER 数据集上的识别准确率提高了 0.96% 和 2.04%。主要原因为基线网络仅在空间上进行特征融合，而

未能捕获帧间的时序信息。这也表明充分利用帧间上下文信息能够有效提升网络的情感识别性能。在嵌入 CA 机制的策略 2～5 中，以不进行关键帧筛选的策略 2 作为基线，仅采用平均池化操作度量帧内情感关联语义与情感整体运动趋势的一致性的策略 3 在两个数据集上的情感识别准确率分别提高了 3.35%、1.28%；策略 4 采用最大池化操作计算帧内情感关联语义与情感显著片段的差异性，在两个数据集上的情感识别准确率分别提高了 3.58%、1.34%；提出的 DPS 策略（策略 5）由于在保留情感显著帧的同时剔除了大量的非峰值帧，因此在两个数据集上的情感识别准确率分别提高了 5.52%、3.70%。类似的，在帧内空间融合子网中嵌入 CrossViT 及 AAM 机制，不同筛选策略间识别性能呈现相同的趋势。这说明了提出的 DPS 策略优于仅采用平均池化或最大池化操作筛选情感显著帧的方式，也说明了在相同的筛选策略下，提出的 AAM 机制优于其他交叉注意力机制。

表 5　不同关键帧筛选策略下 AADPS 网络的情感识别性能

Tab.5　Emotion recognition performance of the AADPS network under different key-frame screening strategies

消融策略	注意力机制	筛选策略		准确率 /%	
		平均池化	最大池化	FABO	CAER
基线网络				82.73	75.20
策略 1				83.69	77.24
策略 2	CA [29]			83.71	76.67
策略 3		√		87.06	77.95
策略 4			√	87.29	78.01
策略 5（DPS）		√	√	89.23	80.37
策略 6	CrossViT [30]			83.86	76.82
策略 7		√		87.42	78.21
策略 8			√	87.51	78.35
策略 9（DPS）		√	√	89.57	80.43
策略 10	AAM			89.28	84.95
策略 11		√		92.88	86.31
策略 12			√	93.23	86.97
策略 13（DPS）		√	√	94.95	88.69

与相关策略相比，本文提出的 AADPS 网络通过嵌入非对称注意力机制和引入双池化关键帧筛选策略，在 FABO 和 CAER 数据集上分别达到了 94.95% 和 88.69% 的识别准确率，较基线网络提高了 12.22% 和 13.49%。这说明提出的 AADPS 网络一方面通过构建非对称注意力机制和两级引导结构提取帧内情感关联语义，增强了面部表情和身姿语不同线索的互补性以及抗干扰能力；另一方面，采用双池化关键帧筛选策略捕获帧间情感显著语义，抑制了非峰值帧的信息冗余。

2.4　AADPS 网络计算性能分析

为说明 AADPS 网络的计算性能，分别统计了不同注意力机制（CA、CrossViT、AAM）及不同网络结构的参数量、浮点运算次数（FLOP）以及在两个数据集上的训练时间、测试时间及识别性能，如表 6 所示（其中空白表示该模块未进行单独训练和测试）。由表 6 可知，与采用 ResNet-50+DSTA-Net 的基线网络相比，基于时空一致性捕获的 3D-CNN+DSTA-Net 网络在 FABO 和 CAER 数据集上的识别准确率仅提高了 0.96%、2.04%，但其参数量和 FLOP 分别了增长了 16.23E6、22079.96E6，且在 FABO 和 CAER 数据集上的训练时长和测试时长分别增加了 0.73 h、5.1 h 和 0.39 min、0.64 min。综合考虑识别性能及计算性能，本文采用 ResNet-50+DSTA-Net 作为基线网络。与基线网络+CA+DPS 以及基线网络 +CrossViT+DPS 的方式相比，AADPS 网络在两个数据集上的情感识别准确率分别提高了 5.72%、8.32% 和 5.38%、8.26%，但其参数量、FLOP 以及在两个数据集上的计算开销增长较少。此外，在测试性能方面，AADPS 网络在两个数据集的测试时长分别小于 3 min 和 30 min，各视频段的平均处理时间均小于 0.8 s，这说明提出的 AADPS 网络能够满足复杂场景下的实时性能要求。

2.5　超参数对网络的性能影响

为说明双池化关键帧筛选策略中关键帧数量 n 对 AADPS 网络识别性能的影响，以 3 为间隔分别统计了提出的 AADPS 网络在 $n \in [3,24]$ 间的情感识别准确率，如图 4 所示。在图 4 中，当 $n<6$ 时，筛选的关键帧数量较少，捕获的帧间时空情感特征缺乏较为完整的时空上下文信息，其情感识别率偏低；当 $n \in [6,12]$ 时，随着双池化关键帧筛选策略保留的显著帧数量逐渐增加，情感识别准确率呈上升趋势；当 $n=12$ 时，筛选的关键帧序列能够较好地表征视频情感片段的动态变化，在两个数据集上的识别准确率均达到了峰值；而随着 n 的持续增大，将在筛选的情感显著帧中引入大量非峰值帧的冗余信息，从而导致情感识别准确率呈下降趋势。

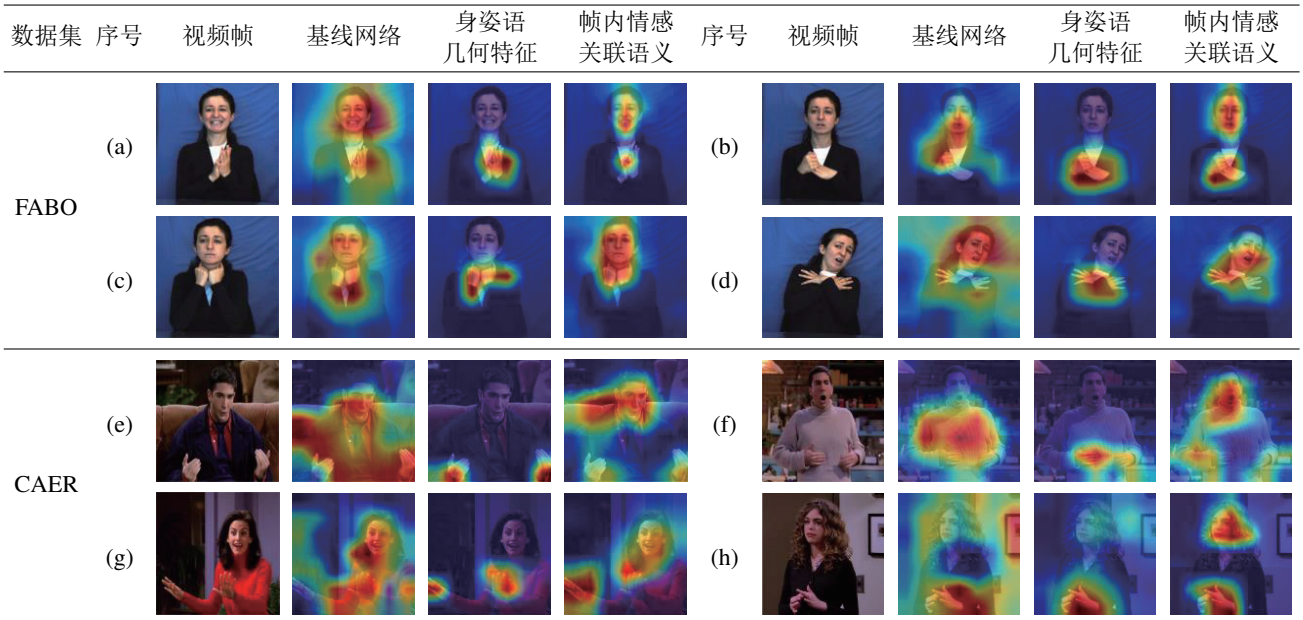
表 6 AADPS 网络的计算及识别性能

Tab.6 Computation and recognition performance of AADPS network

模块 / 网络	参数量 /10^6	FLOP /10^6	训练时长 /h		测试时长 /min		准确率 /%	
			FABO	CAER	FABO	CAER	FABO	CAER
CA [29]	2.13	423.82						
CrossViT [30]	2.34	435.25						
AAM	2.61	443.19						
DPS	2.93	287.34						
3D-CNN+DSTA-Net	58.16	561,317.41	7.89	74.35	2.92	27.35	83.69	77.24
基线网络（ResNet-50+DSTA-Net）	41.93	539,237.45	7.16	69.25	2.53	26.71	82.73	75.20
基线网络+CA+DPS	57.19	558,306.78	7.65	72.01	2.81	26.82	89.23	80.37
基线网络+CrossViT+DPS	57.21	558,357.94	7.69	72.23	2.85	26.89	89.57	80.43
AADPS（本文）	57.82	560,667.03	7.83	73.24	2.90	26.97	94.95	88.69

表 7 非对称注意力机制引导效果的可视化表示

Tab.7 Visualization of the guidance effect of the asymmetric attention mechanism



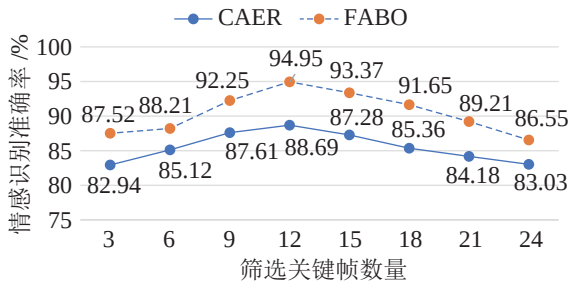


图 4 关键帧数量对情感识别性能的影响

Fig.4 Effect of the number of key frames on emotion recognition performance

2.6 可视化效果分析

为说明非对称注意力机制的引导效果及双池化关键帧筛选策略的筛选效果，分别对其输出的特征图及视频的情感显著度进行可视化处理，如表 7 和图 5 所示。表 7 中各热力图激活特征来源如下：基线网络为 ResNet-50+DSTA-Net 输出的级联特征；“身姿语几何特征”为手势引导姿态的输出特征；“帧内情感关联语义”为面部表情引导身姿语生成的语义。由表 7 可见，基线网络只能粗略关注到面部表情区域，难以聚焦于面部细节和手势部位。在表 7(a) 中，身姿语几何特征能够关注到手势部位，而帧内情感关联语义进一步增加了对张开嘴巴的关注，从而获得了更多 Happiness 信息；与基线热力图相比，表 7(b) 能够聚焦于紧握的双手和面部；此外，表 7(c)(d) 虽面部和手部相距较近，但依然可以聚焦到因 Anxiety 或 Fear 紧闭或张开的嘴巴，同时也对手势部位进行了较大关注。特别地，非对称注意力机制不仅关注到面部表情和手势信息，而

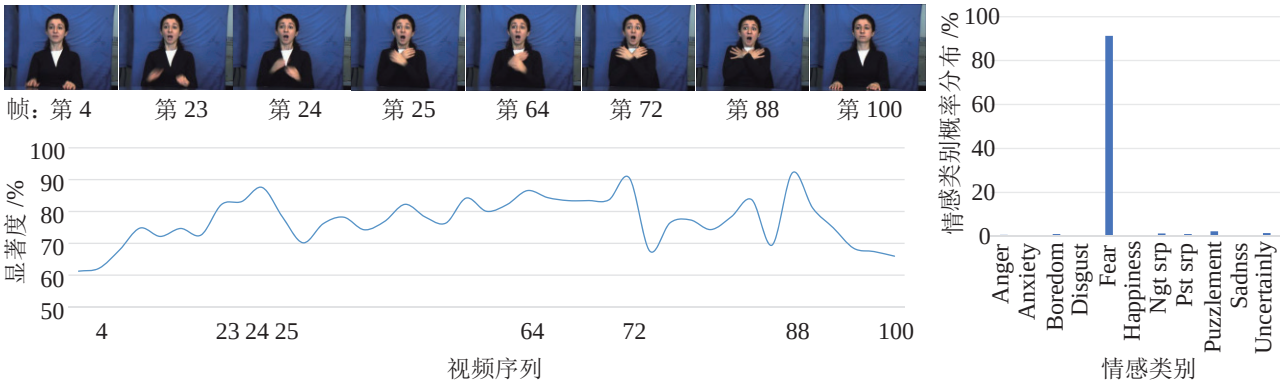


图 5　双池化关键帧筛选策略的可视化表示

Fig.5　Visualization of the dual pooling key-frame screening strategy

且能够抑制冗余信息的干扰。如表 7(e)(f)(g) 中，在关注到面部表情和手势部位的同时，还排除了缺乏情感信息的衣服这一因素的干扰；类似地，表 7(h) 热力图显示其剔除了右上角相框的影响。上述可视化结果表明，由非对称注意力机制获得的身姿语几何特征及帧内情感关联语义，不仅突出了面部、手势等情感显著区域的作用，并且捕获了面部表情和身姿语两者之间的情感相关性。

人类在情感表达过程中，往往呈现出“平静－激活－峰值－减弱－平静”的动态变化模式。因此，视频段的起始与结束阶段将包含大量较低情感强度的平静帧。为了说明提出的双池化关键帧筛选策略的有效性，将 Fear 情感片段的筛选效果进行可视化处理，如图 5 所示。在图 5 中，折线图的横坐标和纵坐标分别表示输入视频序列及对应的情感显著度，柱状图横坐标和纵坐标分别表示不同情感类别及其分类概率大小。从图 5 可以看出，Fear 情感激活阶段、峰值片段以及峰值帧分别标注为 23～25 帧、64～72 帧以及第 88 帧；由双池化关键帧筛选策略度量的各帧情感显著度幅值与标注结果基本保持一致，其中 23～25 帧、64～72 帧和 88 帧的情感显著度均在 80% 以上，且作为情感激活阶段的 23～25 帧的情感显著度远低于作为关键帧的 72、88 帧。此外，图 5 右边柱状图显示该网络能够正确地将输入视频判别为 Fear 且识别准确率高达 94.30%。这表明提出的双池化关键帧筛选策略能够聚焦情感显著帧，从而抑制非关键帧冗余信息的干扰。

2.7　相关方法比较

为说明本文 AADPS 网络的有效性，进一步在 FABO 数据集上与 CNN+LSTM [31]、文 [22] 方法、文 [32] 方法、文 [33] 方法、BDFEI [34]、CMEFA [35] 等方法进行比较，具体结果如表 8 所示。

表 8　基于 FABO 数据集的相关方法比较

Tab.8　Comparison of related methods based on FABO dataset

方法	提出年份	准确率 /%
CNN+LSTM [31]	2021	94.41
文 [22] 方法	2021	93.40
文 [32] 方法	2022	84.30
文 [33] 方法	2023	94.79
BDFEI [34]	2024	93.09
CMEFA [35]	2024	93.58
AADPS（本文）	2024	94.95

在表 8 中，与相关方法相比，本文的 AADPS 网络在 FABO 数据集取得了最佳的情感识别准确率。与仅使用面部表情作为输入的文 [32] 方法相比，AADPS 网络保留了面部表情、姿态和手势 3 类情感细节特征，其准确率提升了 10.65%。与直接对表情和姿态时空特征进行特征级融合的 CNN+LSTM [31]、BDFEI [34] 方法相比，AADPS 网络嵌入非对称注意力机制发挥了面部表情和身姿语不同情感线索间的互补优势，其准确率分别提升了 0.54%、1.86%。与采用整个视频片段进行面部表情和姿态融合的文 [22] 方法以及关注表情－姿态整体时空一致性的文 [33] 方法相比，AADPS 网络引入双池化关键帧筛选策略，其准确率分别提升了 1.55% 和 0.16%。这说明提出的筛选策略能够减小情感动态变化中平静阶段和激活阶段非关键帧带来的干扰。与基于宽－深融合网络的多模态情绪特征耦合网络 CMEFA [35] 相比，AADPS 网络的情感识别准确率提升了 1.37%。这表明本文提出的帧内分层融合方式以及帧间双池化关键帧筛选策略，进一步增强了面部表情和身姿语时空特征的可辨别度。

同时，为了说明 AADPS 网络的泛化能力，本文进一步在大型真实场景情感数据集 CAER 上与

表 9 基于 CAER 数据集的相关方法比较

Tab.9 Comparison of related methods based on CAER dataset

方法	架构	网络	提出年份	准确率 /%
基于单模态的情感识别方法	GCN	CAER-GCN [10]	2023	81.66
	GCN+Transformer	Graph-Tran [10]	2023	77.06
基于多模态融合的情感识别方法	CNN	CAER-Net [28]	2019	77.04
		Zhang [37]	2023	80.10
		MSCNN [36]	2024	81.20
	CNN+GCN	GRERN [20]	2021	86.73
		CMGCN [16]	2022	87.23
	CNN+Transformer	CAHFW-Net [38]	2023	83.76
		AADPS（本文）	2024	88.69

单模态方法以及基于不同架构的多模态方法进行比较，如表 9 所示。在表 9 单模态方法中，基于 GCN（图卷积网络）提取面部表情特征的 CAER-GCN [10] 以及结合 GCN 与 Transformer 捕获面部几何特征的 Graph-Tran [10] 的情感识别准确率分别达到 81.66% 和 77.06%，但普遍低于多模态融合的方法。这表明融合面部表情和身姿语的多模态方法能够充分利用两类线索间的互补信息，提高情感识别的有效性。在多模态融合方法中，MSCNN 网络 [36] 采用多流 CNN 结构融合 3 维人脸动态信息和光流信息，取得了 81.20% 的情感识别准确率；CMGCN 网络 [16] 在采用 CNN 捕获情感信息的同时引入 GCN 增强数据拟合能力，识别准确率进一步提升至 87.23%。与上述采用视频流作为输入的网络相比，Zhang [37] 构建基于弱监督的跨模态时间擦除网络，情感识别准确率达到 80.10%；CAHFW-Net 网络 [38] 通过构建特征融合模块捕获表情与姿态间的相关性，准确率达到了 83.76%；总的来说，由于 GCN 和 Transformer 具有强大的时序捕获能力，以上基于 CNN+GCN/Transformer 的网络 [16,20,38]，其情感识别精度普遍优于其他网络；但相关方法缺乏多模态的相关性建模，且忽略了非关键帧信息冗余的影响，导致情感识别性能均不及提出的 AADPS 网络。提出的 AADPS 网络以面部纹理流、姿态和手势的骨骼流为输入，采用 1 级非对称注意力机制提取身姿语几何特征，在获取手势与姿态情感相关性的同时充分保留了手势线索的细节语义；此外，利用 2 级非对称注意力机制获取帧内情感关联语义，进一步提升了帧内的面部表情和身姿语不同线索的互补性；同时，引入双池化关键帧筛选策略捕获帧间情感显著语义，既保留了优质情感显著帧，又剔除大量的非峰值帧的影响，其情感识别准确率达到 88.69%。

3 总结（Conclusion）

为提升复杂场景下的情感识别性能，提出一种基于非对称注意与双池化筛选的情感识别网络。在帧内空间融合子网中，构建非对称注意力机制及两级引导结构捕获面部表情和身姿语的情感相关性，从而凸显不同情感线索的重要性。在帧间时空融合子网中，设计双池化关键帧筛选策略度量各帧情感显著度并筛选情感显著帧，以此抑制非峰值帧的信息冗余。在 FABO 和 CAER 视频数据集上评价了所提出网络的情感类别 F1 值、平均 F1 值、准确率等指标，统计了嵌入模块的参数量、FLOP 等计算指标，并讨论了筛选关键帧数量对情感识别准确率的影响；同时，设计了不同的消融策略，说明了提出的非对称注意机制以及双池化关键帧筛选策略的有效性，并对其输出的特征图及视频情感显著度进行了可视化。实验结果表明，提出的网络在 FABO 和 CAER 数据集上的情感识别准确率分别达到了 94.95% 和 88.69%，较基线网络分别提升了 12.22% 和 13.49%；与相关方法相比，提出的方法在复杂场景下具有更好的情感识别性能。然而，在复杂场景中，用户的位姿、场景的光照及遮挡等因素不可避免地导致面部或姿态信息缺失。如何将信息补偿机制引入至多模态情感识别模型中将是下一步的研究工作。

参考文献（References）

[1] WORTMAN B, WANG J Z. HICEM: A high-coverage emotion model for artificial emotional intelligence[J]. IEEE Transactions on Affective Computing, 2024, 15(3): 1136-1152.

[2] 黄忠，任福继，胡敏，等. 基于 Transformer 架构和 B 样条平滑约束的机器人面部情感迁移网络[J]. 机器人，2023，45(4)：395-408.
HUANG Z, REN F J, HU M, et al. Robotic facial emotion transfer network based on transformer framework and B-spline smoothing constraint[J]. Robot, 2023, 45(4): 395-408.

[3] MARTINEZ L, FALVELLO V B, AVIEZER H, et al. Contributions of facial expressions and body language to the rapid perception of dynamic emotions[J]. Cognition and Emotion, 2016, 30(5): 939-952.

[4] 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. DOI: 10.1016/j.cosrev.2023.100545.

[5] 祝文斌，苑晶，朱书豪，等. 低光照场景下基于序列增强的移动机器人人体检测与姿态识别[J]. 机器人，2022，44(3)：299-309.
ZHU W B, YUAN J, ZHU S H, et al. Sequence-enhancement-based human detection and posture recognition of mobile robots in low illumination scenes[J]. Robot, 2022, 44(3): 299-309.

[6] 杨天昊，李云开，王雅昕，等. 服务机器人的触觉手势识别[J]. 机器人，2022，44(3)：310-320.
YANG T H, LI Y K, WANG Y X, et al. Touch gesture recognition for service robots[J]. Robot, 2022, 44(3): 310-320.

[7] WU S H, 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.

[8] 陶建华，范存航，连政，等. 多模态情感识别与理解发展现状及趋势[J]. 中国图象图形学报，2024，29(6)：1607-1627.
TAO J H, FAN C H, LIAN Z, et al. Development of multimodal sentiment recognition and understanding[J]. Journal of Image and Graphics, 2024, 29(6): 1607-1627.

[9] CHAKRAVARTHI S S, RAO B N K, CHALLA N P, et al. Gesture recognition for enhancing human computer interaction[J]. Journal of Scientific & Industrial Research, 2023, 82(4): 438-443.

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

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

[12] LIU X, LI S, WANG M. Hierarchical attention-based multimodal fusion network for video emotion recognition[J]. Computational Intelligence and Neuroscience, 2021. DOI: 10.1155/2021/558504.

[13] 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.

[14] PALASH M, BHARGAVA B. EMERSK – Explainable multimodal emotion recognition with situational knowledge[J]. IEEE Transactions on Multimedia, 2024, 26: 2785-2794.

[15] ZHOU S, WU X, 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): 1400-1423.

[16] 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.

[17] HAN X, CHEN F Y, BAN J R. FMFN: A fuzzy multimodal fusion network for emotion recognition in ensemble conducting [J]. IEEE Transactions on Fuzzy Systems, 2025, 33(1): 168-179.

[18] GU Y, ZHANG X, HUANG J Y, et al. WiFE: Wifi and vision based unobtrusive emotion recognition via gesture and facial expression[J]. IEEE Transactions on Affective Computing, 2023, 14(4): 2567-2581.

[19] 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: 103679-103685.

[20] GAO Q Q, ZENG H X, LI G, et al. Graph reasoning-based emotion recognition network[J]. IEEE Access, 2021, 9: 6488-6497.

[21] 李博凯，吴从中，项柏杨，等. 微表情峰值帧定位引导的分类算法[J]. 中国图象图形学报，2024，29(5)：1447-1459.
LI B K, WU C Z, XIANG B Y, et al. Apex frame spotting and recognition of micro-expression by optical flow[J]. Journal of Image and Graphics, 2024, 29(5): 1447-1459.

[22] 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.

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

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

[25] HE K M, ZHANG X Y, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 770-778,

[26] XU M Y, GUO L M, WU H. Robust abnormal human-posture recognition using OpenPose and multiview cross-information [J]. IEEE Sensors Journal, 2023, 23(11): 12370-12379.

[27] SHI L, ZHANG Y, CHENG J, et al. Decoupled spatial-temporal attention network for skeleton-based action recognition[DB/OL]. (2020-07-07) [2023-05-06]. https://arxiv.org/abs/2007.03263.

[28] LEE J, LIM S, LIM S, et al. Context-aware emotion recognition networks[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2020: 10142-10151.

[29] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[DB/OL]. (2023-08-02) [2024-09-01]. https://arxiv.org/abs/1706.03762.

[30] CHENC F R, FAN Q, PANDA R. CrossViT: Cross-attention multi-scale vision transformer for image classification[C]// IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2021: 357-366.

[31] AQDUS C M, NUNES R, NASROLLAHI K, et al. Deep emotion recognition through upper body movements and facial expression[C]//16th International Conference on Computer Vision Theory and Applications. Piscataway, USA: IEEE, 2021: 669-679.

[32] 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.

（下转第 591 页）

[9] JIANG H C, XU Y M, LI K J, et al. RoDyn-SLAM: Robust dynamic dense RGB-D SLAM with neural radiance fields[J]. IEEE Robotics and Automation Letters, 2024, 9(9): 7509-7516.
[10] LI H A, MENG X Q, ZUO X X, et al. PG-SLAM: Photo-realistic and geometry-aware RGB-D SLAM in dynamic environments[J]. IEEE Transactions on Robotics, 2025, 41: 6084-6101.
[11] BOLYA D, ZHOU C, XIAO F Y, et al. YOLACT: Real-time instance segmentation[C]//IEEE/CVF International Conference on Computer Vision. Piscataway, USA: IEEE, 2019: 9157-9166.
[12] KUMAR A, VASHISHTHA G, GANDHI C P, et al. Novel convolutional neural network (NCNN) for the diagnosis of bearing defects in rotary machinery[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70. DOI: 10.1109/TIM.2021.3055802.
[13] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2016: 21-37.
[14] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2017: 2961-2969.
[15] 张涛. 无参考图像模糊度评价方法研究[D]. 大连：大连海事大学，2015.
ZHANG T. Research on evaluation method of blurriness of non-reference image[D]. Dalian: Dalian Maritime University, 2015.
[16] HARTLEY R, ZISSERMAN A. Multiple view geometry in computer vision[M]. Cambridge, UK: Cambridge University Press, 2003.
[17] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state-of-the-art superpixel methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282.
[18] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2012: 573-580.
[19] LIU Y B, MIURA J. RDS-SLAM: Real-time dynamic SLAM using semantic segmentation methods[J]. IEEE Access, 2021, 9: 23772-23785.
[20] YU C, LIU Z X, LIU X J, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 1168-1174.
[21] MIN F, WU Z, LI D, et al. COEB-SLAM: A robust VSLAM in dynamic environments combined object detection, epipolar geometry constraint, and blur filtering[J]. IEEE Sensors Journal, 2023, 23(21): 26279-26291.
[22] PALAZZOLO E, BEHLEY J, LOTTES P, et al. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2019: 7855-7862.

作者简介：

党博文（1999 –）男，硕士生，研究领域：视觉 SLAM。
熊根良（1978 –）男，博士，教授，研究领域：机器人技术，机器人柔顺控制，人－机器人交互，医疗服务机器人。

（上接第 517 页）

[33] XIE B J, PARK C H. Can you guess my moves? Playing charades with a humanoid robot employing mutual learning with emotional intelligence[C]//ACM/IEEE International Conference on Human-Robot Interaction. New York, USA: ACM, 2023: 667-671.
[34] 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: 1-7.
[35] CHEN L, 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.
[36] 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(6A). DOI：10.11896/jsjkx.230700210.
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(6A). DOI: 10.11896/jsjkx.230700210.
[37] ZHANG Z, WANG L, YANG J. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023: 18888-18897.
[38] ZHOU S, WU X, 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): 1400-1423.

作者简介：

黄　忠（1981 –），男，博士，教授。研究领域：类人机器人，自然人机交互。