基于关系感知和标签消歧的细粒度面部表情识别算法

刘雅芝[1,2],许喆铭[1,2*],郎丛妍[1,2],王 涛[1,2],李沁东[1,2]
(1. 北京交通大学计算机科学与技术学院,北京 100044;2. 北京交通大学交通大数据与人工智能教育部重点实验室,北京 100044)

摘 要: 细粒度表情识别任务因其包含更丰富真实的人类情感而备受关注. 现有面部表情识别算法通过提取局部关键区域等方式学习更优的图像表征. 然而,这些方法忽略了图像数据集内在的结构关系,且没有充分利用标签间的语义关联度以及图像和标签间的相关性,导致所学特征带来的性能提升有限. 其次,现有细粒度表情识别方法并未有效利用和挖掘粗细粒度的层级关系,因而限制了模型的识别性能. 此外,现有细粒度表情识别算法忽略了由于标注主观性和情感复杂性导致的标签歧义性问题,极大影响了模型的识别性能. 针对上述问题,本文提出一种基于关系感知和标签消歧的细粒度面部表情识别算法(fine-grained facial expression recognition algorithm based on Relationship-Awareness and Label Disambiguation,RALD). 该算法通过构建层级感知的图像特征增强网络,充分挖掘图像之间、层级标签之间以及图像和标签之间的依赖关系,以获得更具辨别性的图像特征. 针对标签歧义性问题,算法设计了基于近邻样本的标签分布学习模块,通过整合邻域信息进行标签消歧,进一步提升模型识别性能. 在细粒度表情识别数据集 FG-Emotions 上算法的准确度达到 97.34%,在粗粒度表情识别数据集 RAF-DB 上比现有主流表情分类方法提高了 0.80% ~ 4.55%.

关键词: 细粒度面部表情识别;注意力机制;关系感知;特征优化;标签分布学习
基金项目: 国家自然科学基金(No.62072027,No.62376020)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3336-11
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240364

Fine-Grained Facial Expression Recognition Algorithm Based on Relationship-Awareness and Label Disambiguation

LIU Ya-zhi[1,2], XU Zhe-ming[1,2*], LANG Cong-yan[1,2], WANG Tao[1,2], LI Yi-dong[1,2]
(1. *School of Computer Science & Technology*, *Beijing Jiaotong University*, *Beijing* 100044, *China*;
2. *Key Laboratory of Big Data and Artificial Intelligence in Transportation* (*Beijing Jiaotong University*), *Ministry of Education*, *Beijing* 100044, *China*)

Abstract: There has been a growing interest in fine-grained facial expression recognition due to its ability to capture more subtle and realistic human emotions. Existing facial expression recognition algorithms enhance image representations by extracting local key regions and other relevant features. However, these methods disregard the inherent structural relationships within the image dataset and fail to fully exploit the semantic correlation between labels and the relationship between images and labels, which restricts the enhancement of feature learning. Besides, current fine-grained expression recognition methods do not effectively explore and utilize the hierarchical relationship between coarse and fine-grained levels, which limits the recognition performance of the model. In addition, existing fine-grained expression recognition algorithms ignore the label ambiguity problem caused by labeling subjectivity and emotional complexity, which greatly affects the recognition performance of the model. To address these issues, we propose a fine-grained facial expression recognition algorithm based on relationship-awareness and label disambiguation (RALD). This algorithm enhances image features by constructing a hierarchy-aware image feature enhancement network, thoroughly exploring the dependencies among images, hierarchical labels, and between images and labels to obtain more discriminative image features. As for the issue of label ambiguity, this algorithm designs a nearest neighbors-based label

收稿日期:2024-04-22;修回日期:2024-08-11;责任编辑:覃怀银
*通讯作者:许喆铭

distribution learning module, which further improves recognition performance by integrating neighborhood information for label disambiguation. Our algorithm achieves 97.34% in terms of accuracy on the FG-Emotions dataset for fine-grained expression recognition. Additionally, it outperforms existing mainstream facial expression recognition algorithms by 0.80% to 4.55% on the RAF-DB dataset for coarse-grained expression recognition.

Key words: fine-grained facial expression recognition; attention mechanism; relation awareness; feature optimization; label distribution learning

Foundation Item(s): National Natural Science Foundation of China (No.62072027, No.62376020)

1 引言

近年来,面部表情识别(Facial Expression Recognition,FER)逐渐成为情感计算领域和计算机视觉中的研究热点,旨在将人类静态面部图像中的情感转化为计算机可理解的形式,以实现更智能、更人性化的交互. 目前,该任务已被广泛应用于人机交互[1,2]、课堂教学[3]和卫生保健[4]等领域. 然而,现阶段的面部表情识别任务主要基于Ekman[5]提出的六类(或加入中性表情的七类)基本情绪,忽略了人类情绪和面部表情的丰富度和多样性,限制了情感计算中的细粒度语义理解. 因此,研究细粒度面部表情识别任务具有十分重要的现实意义和应用价值.

传统的基于基本表情的面部表情识别算法主要致力于提取更具辨别性的特征表示,具体方法包括构建更强的网络结构[6,7]、获取局部关键区域特征[8~10]以及引入额外信息如身份信息[11,12]等,但细粒度表情相对于基本表情有着更加明显的类间差异小、类内差异大的特点. 基于基本表情识别的算法大多关注图像全局或图像内的局部区域特征,难以适用于细粒度表情识别任务. 同时,这些算法忽略了图像数据集的内在拓扑结构关系,也没有充分利用标签之间的语义关联以及图像与标签之间的相关性,而标签语义关系的作用在视觉问答和图像字幕生成等任务中已经得到证明,其通过不同模态信息之间的互补性,提升模型对图像关系的理解能力. 因此,如何有效地挖掘和利用图像间、标签间以及图像与标签之间的关联性,以进一步优化图像特征表示,成为面部表情识别的一大挑战.

在细粒度表情识别任务中天然存在粗细粒度标签的层级关系. 因此,如何充分利用这一层级结构获得更鲁棒的图像特征表示是细粒度表情识别任务中的另一大挑战. 此外,细粒度面部表情识别的标注涉及到标注者对情感的理解,而不同文化背景、个人经历等因素都可能影响标注者对表情的理解和分类[13],同时细粒度情感表征本身的复杂性和多样性也会导致较大的标签歧义性,进而降低了模型的识别性能. 虽然Zhu等人[14]提出的两阶段的关系挖掘网络R3HO-Net模型通过k近邻算法和标签从属关系构建图结构,并用图卷积神经网络更新关系信息. 然而,该方法对参数敏感,且未充分利用标签层级关系,同时也未考虑标签歧义性对模型性能的限制.

针对细粒度面部表情识别中的难点以及现有方法的不足,本文提出了一种基于关系感知和标签消歧的细粒度面部表情识别算法(fine-grained facial expression recognition algorithm based on Relationship-Awareness and Label Disambiguation,RALD). 它主要由两个部分组成,分别是层级感知的图像特征增强网络以及基于近邻样本的标签分布学习模块. 其中,层级感知的图像特征增强网络包含三个模块,即图像拓扑关系表示、层级感知的标签关系表示以及标签引导的图像特征优化模块,通过对图像间、标签间及图像与标签之间的关系进行建模,以生成更有辨别力的图像特征表示;基于近邻样本的标签分布学习模块利用近邻信息生成标签分布,并约束其与真实标签分布的相似性,缓解了单一标签带来的标签歧义性问题.

2 相关工作

2.1 细粒度面部表情识别

面部表情识别的研究可追溯到由Ekman[5]提出的六类基本情绪理论,包括惊讶(surprise)、厌恶(disgust)、恐惧(fear)、快乐(happiness)、悲伤(sadness)和愤怒(anger). 随后,基于这一经典心理学理论,基于基本情绪的粗粒度表情识别大致可以分为基于特征提取的方法[11,12,15~22]和基于损失函数设计的方法[23,24].

然而,粗粒度的基本表情并不能充分体现表情的复杂性和微妙性. 为了涵盖丰富的人脸表情,如图1所示,Liang等人[25]构建了一个包含33种表情类别的自然环境下的细粒度表情数据集FG-Emotions,并提出了一种基于多尺度动作单元的MSAU-Net模型,通过检测和“放大”面部动作单元来定位最有辨别性的面部区域. Zhu等人[14]提出了一种基于图卷积神经网络的R3HO-Net模型,利用关系推理和分层关系优化将细粒度表情识别问题建模为关系映射问题.

2.2 面部表情识别中的注意力机制

注意力机制[26]由于其强大的特征提取和关键信息筛选能力[27,28]被广泛地应用于细粒度识别任务中. 视觉任务中. 其在人脸表情识别上有许多应用,例如TransFER模型[20]、APViT模型[29]和MRAN模型[18]通过注意力机制获取最相关的面部动作单元区域,从而学习具有辨

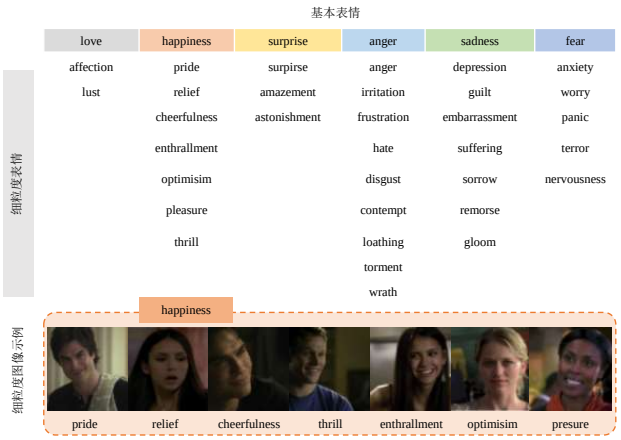


图1 FG-Emotions的层次结构及图像示例

别性的表情特征. 然而,细粒度面部表情识别中,图像之间具有较高的视觉相似性,仅通过挖掘图像内部的细微差异带来的性能提升有限. 因此,本文探索利用自注意力和交叉注意力机制,充分挖掘图像与标签以及图像之间的相关性. 一方面,利用标签语义以及层级结构引导模型提取更优的特征表示;另一方面利用图像近邻相关性缓解数据中存在的标签歧义性问题.

2.3 标签分布学习

标签分布学习最初由Geng[30]于2016年提出,旨在通过学习图像的标签分布来替代单一标签信息,以此更好地解决标签歧义性问题[31]. 在表情识别任务中,由于标签分布标注较为困难且工作量庞大,目前大多数FER数据集难以为每个样本提供情绪标签分布信息. 因此,现有研究转向通过学习表情的标签分布来监督单标签分类问题. 例如,Le等人[32]和Chen等人[33]利用辅助任务得到的信息来构建标签分布以监督模型训练. Zhao等人[34]通过训练标签分布生成器直接生成标签分布. 然而,上述方法均需要借助额外的标注或者辅助信息,致使计算复杂度和成本较高. 不同于上述方法,本文利用数据自身的特征空间信息,有效挖掘图像拓扑关系,并利用近邻信息构建标签分布,实现标签信息增强.

3 RALD模型

如图2为本文所提出的RALD模型的整体框架,其主要由图像及标签语义编码器、层级感知的图像特征增强网络(Hierarchy-aware Image Feature Enhancement Network, HIFE-Net)以及基于近邻样本的标签分布学习模块(Nearest Neighbors-based Label Distribution Learning, NNLDL)组成. 其中,层级感知的图像特征增强网络包含图像拓扑关系表示(Image Topological Relation Representation, ITRR)、层级感知的标签关系表示(Hierarchy-aware Label Relation Representation, HLRR)以及标签引导的图像特征优化模块(Label Guided Image Feature Optimization Module, LGIFO). 具体地,本文首先通过图像编码器和标签语义编码器,提取图像特征表示和标签语义信息;之后,通过层级感知的图像特征增强网络,同时有效感知图像之间、标签之间以及图像和标签之间的依赖关系,以对图像特征进行优化并获得更鲁棒的图像表征;此外,通过基于近邻样本的标签分布学习模块构造样本的伪预测标签分布,并用于监督模型预测的样本标签分布,以此缓解标签歧义性问题对模型的影响.

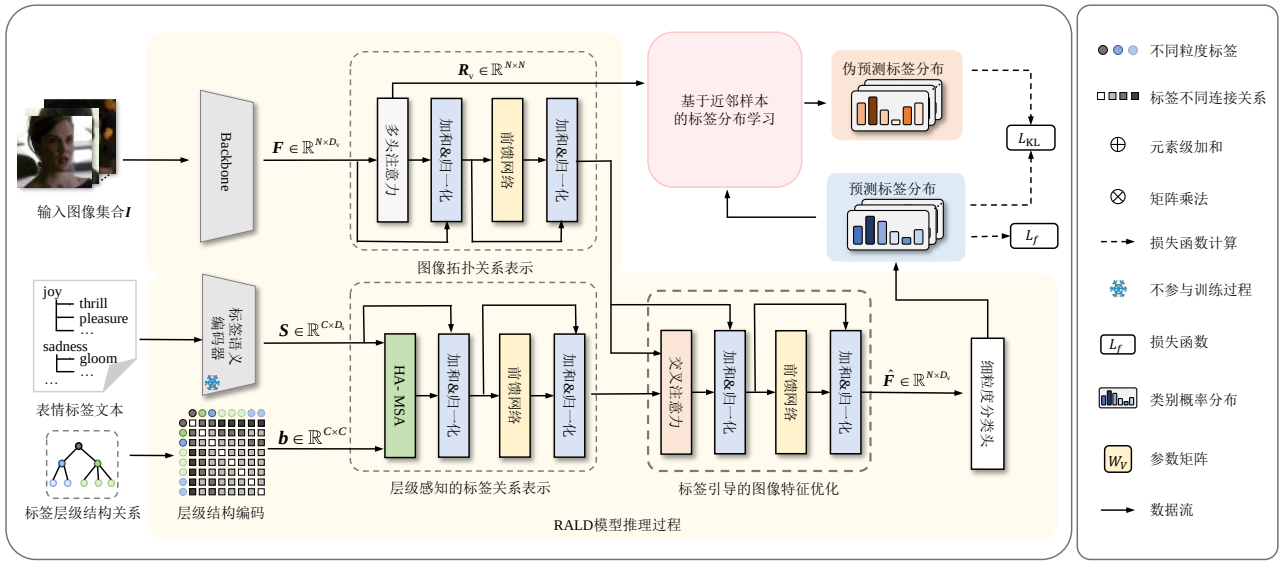


图2 RALD模型整体框架

3.1 问题定义

在细粒度面部表情识别任务中,粗细粒度标签之间存在天然的层级结构. 本文首先将这一层级结构建模为一个无向无权图 G,图 G 的结点集 $V=\{v_i|i=1,\cdots,C\}$ 代表粗细粒度标签组成的集合,其中,C 表示粗细粒度标签类别的总数目,图 G 的边集 E 代表相邻层级标签之间的边的集合. 因此,细粒度表情识别任务可以形式化为:已知多粒度标签层级结构 G,给定一个批次的人脸面部表情图像 $\boldsymbol{I}$,$\boldsymbol{I}_i$ 表示第 i 个表情图像. 细粒度面部表情识别旨在输出这些图像对应的识别结果 $\hat{\boldsymbol{Y}}=(\hat{y}_0,\hat{y}_1,\cdots,\hat{y}_{N-1})$ 其中,N 代表一个批次的图像数量,$\hat{y}_i\in[0,C_f-1]$ 为第 i 个图像预测的类别标签,C_f 代表细粒度表情类别标签的总数目.

3.2 图像与标签语义编码器

(1)图像编码器. 如图 2 所示,给定一批次图像 $\boldsymbol{I}$,图像编码器 F_{v} 将输出图像的特征表示. 如式(1)所示:

$$\boldsymbol{F}=F_{\mathrm{v}}(\boldsymbol{I}) \tag{1}$$

其中,$\boldsymbol{F}\in\mathbb{R}^{N\times D_{\mathrm{v}}}$,$\boldsymbol{f}_i$ 是第 i 个图像的特征向量,D_{v} 表示 $\boldsymbol{f}_i$ 的向量维度.

(2)标签语义编码器. 标签语义编码器旨在获得标签集 V 中的每一个标签的特征表示. word2vec[35]是一个可以构建词向量的三层神经网络,本文利用预训练的 word2vec 模型将每个类别标签 $v_i\in V$ 映射为 300 维的词向量,相似语义的标签对应词向量的余弦相似度也更高. 需要注意的是,语义特征提取过程不参与后续梯度反向传播. 因此,给定包含多层级的标签集 V,标签语义特征 $\boldsymbol{S}\in\mathbb{R}^{C\times D_{\mathrm{s}}}$ 提取过程如式(2)所示:

$$\boldsymbol{S}=F_{\mathrm{s}}(V) \tag{2}$$

其中,D_{s} 表示每个类别标签映射后的特征维度.

3.3 层级感知的图像增强网络

3.3.1 图像拓扑关系表示

由图像编码器获得初始图像特征 $\boldsymbol{F}$ 后,本文利用自注意力机制,挖掘图像之间的内在拓扑结构关系. 具体地,首先通过三个线性层分别将图像特征 $\boldsymbol{F}$ 映射到对应的特征空间,并将对应特征分别表示为 $\boldsymbol{Q}_{\mathrm{v}}$、$\boldsymbol{K}_{\mathrm{v}}$ 和 $\boldsymbol{V}_{\mathrm{v}}$ 之后,本文利用 $\boldsymbol{Q}_{\mathrm{v}}$、$\boldsymbol{K}_{\mathrm{v}}$ 和 $\boldsymbol{V}_{\mathrm{v}}$ 分别计算自注意力,并扩展到多头注意力,最后经过求和以及归一化操作,得到更新后的图像拓扑关系特征 $\tilde{\boldsymbol{F}}$. 这一过程如式(3)~(7)所示:

$$\boldsymbol{Q}_{\mathrm{v}}=\boldsymbol{F}\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{K}_{\mathrm{v}}=\boldsymbol{F}\boldsymbol{W}_{\mathrm{K}},\boldsymbol{V}_{\mathrm{v}}=\boldsymbol{F}\boldsymbol{W}_{\mathrm{V}} \tag{3}$$

$$\boldsymbol{R}_{\mathrm{v}}=\lambda\boldsymbol{Q}_{\mathrm{v}}\boldsymbol{K}_{\mathrm{v}}{}^{\mathrm{T}} \tag{4}$$

$$\boldsymbol{F}_{\mathrm{SA}}=\mathrm{softmax}(\boldsymbol{R}_{\mathrm{v}})\boldsymbol{V}_{\mathrm{v}} \tag{5}$$

$$\boldsymbol{F}_{\mathrm{MSA}}=\mathrm{concat}(\boldsymbol{F}_{\mathrm{SA}}^{1},\cdots,\boldsymbol{F}_{\mathrm{SA}}^{h})\boldsymbol{W}_{\mathrm{MSA}} \tag{6}$$

$$\tilde{\boldsymbol{F}}=\mathrm{norm}(\boldsymbol{F}_{\mathrm{MSA}}+\mathrm{FFN}(\boldsymbol{F}_{\mathrm{MSA}})) \tag{7}$$

其中,$\boldsymbol{W}_{\mathrm{Q}}$、$\boldsymbol{W}_{\mathrm{K}}$ 和 $\boldsymbol{W}_{\mathrm{V}}$ 分别表示不同线性层的参数矩阵,$\boldsymbol{R}_{\mathrm{v},ij}$ 表示第 i 个图像和第 j 个图像之间的相关性. λ 表示缩放参数. $\boldsymbol{F}_{\mathrm{SA}}$ 和 $\boldsymbol{F}_{\mathrm{MSA}}$ 分别表示单头和多头注意力模块的输出结果,其中 h 表示注意力头的个数. concat、norm 和 FFN 分别表示特征拼接操作、归一化操作和前馈神经网络计算.

3.3.2 层级感知的标签关系表示

如图 3 所示为层级感知的多头自注意力机制(Hierarchical-Aware Multi-Head Self-Attention, HA-MSA). 不同于图像特征,标签信息包括语义及标签之间的层级结构信息. 其中,标签层级关系作为一种图结构,一种直观的做法是使用图神经网络如图卷积神经网络[14]对其进行建模. 然而,图神经网络相对简单,容易出现过平滑问题. 受 Graphormer[36]的启发,本文在多头注意力模块中,引入层级结构编码(Hierarchical Encoder, HE),从而融入标签层级关系. 类似于式(4),这一过程如式(8)所示:

$$\boldsymbol{R}_{\mathrm{s},ij}=(\boldsymbol{h}_i\overline{\boldsymbol{W}}_{\mathrm{Q}})(\boldsymbol{h}_j\overline{\boldsymbol{W}}_{\mathrm{K}})^{\mathrm{T}}+b_{\xi(v_i,v_j)} \tag{8}$$

其中,$\boldsymbol{R}_{\mathrm{s},ij}$ 代表多头注意力机制中相关性矩阵 $\boldsymbol{R}_{\mathrm{s}}$ 的第 i 行第 j 列元素,表示第 i 个标签与第 j 个标签的依赖程度. $\overline{\boldsymbol{W}}_{\mathrm{Q}}$ 和 $\overline{\boldsymbol{W}}_{\mathrm{K}}$ 代表不同线性层的参数矩阵. λ 为缩放参数. $\boldsymbol{h}_i$ 和 $\boldsymbol{h}_j$ 分别表示第 i 个和第 j 个隐向量特征. $b_{\xi(v_i,v_j)}$ 表示层级感知编码的偏置项.

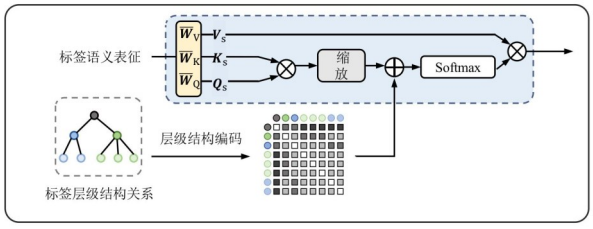


图 3　层级感知的多头自注意力(HA-MSA)

给定一个标签层级结构 G,本文定义一个由 $\boldsymbol{\xi}(v_i,v_j)$ 索引的可学习的标量 $b_{\xi(v_i,v_j)}$ 来表示结构编码结果. 具体地,$\boldsymbol{\xi}(v_i,v_j)\in\mathbb{R}^{C\times C}$ 用于度量 G 中各节点之间的空间结构关系,其表示如式(9)所示:

$$\xi(v_i,v_j)=\begin{cases}\mathrm{SPD}(v_i,v_j), & (v_i,v_j)\in\varepsilon\\ 0, & i=j\\ \sigma, & \text{其他}\end{cases} \tag{9}$$

其中,ε 表示粗粒度标签与细粒度标签对应节点的边的集合. $\mathrm{SPD}(v_i,v_j)$ 表示从节点 v_i 到节点 v_j 的最短路径距离(Shortest Path Distance, SPD). σ 为一个与最短路径距离相关的常数. 由此,可构建一个关于 $\boldsymbol{\xi}(v_i,v_j)$ 的递减函数 $b_{\xi(v_i,v_j)}$,从而将图结构编码为自注意力机制中的偏置项. 当节点 v_i 和 v_j 在空间结构上的 SPD 越远时,其相关性越低,对应的偏置 $b_{\xi(v_i,v_j)}$ 则越小.

通过上述层级结构编码，RALD模型能够有效地利用细粒度表情中的标签结构信息. 之后遵循式(10)~(12)，本文对标签语义信息进行强化，从而获得层级感知的标签关系的标签特征$\tilde{\boldsymbol{S}}$.

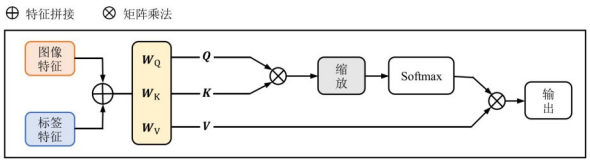
$$\boldsymbol{S}_{\mathrm{SA}}=\mathrm{softmax}(\boldsymbol{R}_{\mathrm{s}})\boldsymbol{V}_{\mathrm{s}} \tag{10}$$

$$\boldsymbol{S}_{\mathrm{MSA}}=\mathrm{concat}(\boldsymbol{S}_{\mathrm{SA}}^{1},\cdots,\boldsymbol{S}_{\mathrm{SA}}^{h})\boldsymbol{W}_{\mathrm{MSA}} \tag{11}$$

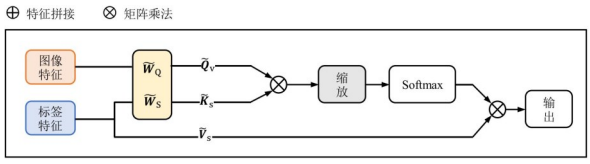
$$\tilde{\boldsymbol{S}}=\mathrm{norm}(\boldsymbol{S}_{\mathrm{MSA}}+\mathrm{FFN}(\boldsymbol{S}_{\mathrm{MSA}})) \tag{12}$$

3.3.3 标签引导的图像特征优化

如图4(*a*)所示，现有面部表情算法往往对图像特征和标签特征进行简单级联，并通过多头注意力模块对融合特征进行建模，忽略了图像特征和标签特征之间的相关性. 针对这一问题，如图4(*b*)，本文提出标签引导的图像特征优化模块，通过交叉注意力机制建模标签与图像之间的关系，利用标签信息对图像信息进行增强.



(*a*) 常见的基于注意力机制的数据融合方法



(*b*) RALD模型的数据融合方法

图4 常见的数据融合方法与本文所提数据融合方法对比

具体地，本文利用交叉注意力模块的多头注意力机制将图像特征$\tilde{\boldsymbol{F}}$和标签特征表示$\tilde{\boldsymbol{S}}$分别映射到不同特征空间，以获得基于图像特征的查询矩阵$\tilde{\boldsymbol{Q}}_{\mathrm{v}}$和基于标签特征的键值矩阵$\tilde{\boldsymbol{K}}_{\mathrm{s}}$和$\tilde{\boldsymbol{V}}_{\mathrm{s}}$. 其具体过程如式(13)所示：

$$\tilde{\boldsymbol{Q}}_{\mathrm{v}}=\tilde{\boldsymbol{F}}\tilde{\boldsymbol{W}}_{\mathrm{Q}},\tilde{\boldsymbol{K}}_{\mathrm{s}}=\tilde{\boldsymbol{S}}\tilde{\boldsymbol{W}}_{\mathrm{K}},\tilde{\boldsymbol{V}}_{\mathrm{s}}=\tilde{\boldsymbol{S}} \tag{13}$$

由此，针对三元组$\left(\tilde{\boldsymbol{Q}}_{\mathrm{v}},\tilde{\boldsymbol{K}}_{\mathrm{s}},\tilde{\boldsymbol{V}}_{\mathrm{s}}\right)$可依据式(4)获得图像与标签之间的相关性矩阵$\boldsymbol{R}_{\mathrm{vs}}$，并根据$\boldsymbol{R}_{\mathrm{vs}}$更新图像的特征表示，以此通过标签信息的引导，实现对图像特征的优化，获得最终的图像特征表示$\hat{\boldsymbol{F}}$. 这一过程如式(14)~(17)所示：

$$\boldsymbol{R}_{\mathrm{vs}}=\lambda\tilde{\boldsymbol{Q}}_{\mathrm{v}}\tilde{\boldsymbol{K}}_{\mathrm{s}}{}^{\mathrm{T}} \tag{14}$$

$$\boldsymbol{X}_{\mathrm{SA}}=\mathrm{softmax}(\boldsymbol{R}_{\mathrm{vs}})\tilde{\boldsymbol{V}}_{\mathrm{s}} \tag{15}$$

$$\boldsymbol{X}_{\mathrm{MSA}}=\mathrm{concat}(\boldsymbol{X}_{\mathrm{SA}}^{1},\cdots,\boldsymbol{X}_{\mathrm{SA}}^{h})\boldsymbol{W}_{\mathrm{MSA}} \tag{16}$$

$$\hat{\boldsymbol{F}}=\mathrm{norm}(\boldsymbol{X}_{\mathrm{MSA}}+\mathrm{FFN}(\boldsymbol{X}_{\mathrm{MSA}})) \tag{17}$$

3.4 基于近邻样本的伪预测标签分布生成

在面部表情数据标注过程中，不同背景的人对面部表情的认知可能不同，使得标注具有较强的主观性. 此外，由于人类情感的复杂性，面部表情通常包含多种情感类别，导致数据存在严重的标签歧义性问题，进而极大限制了模型性能. 然而现有方法往往需借助额外信息，致使计算成本较高. 针对上述问题，本文提出基于近邻样本的标签分布生成模块，探索利用数据集自身的特征空间，基于“样本的相似特征近邻应该具有相似情绪”的假设，通过加权融合邻居样本的标签信息构造样本标签分布.

具体地，本文首先通过图像拓扑关系表示模块获取图像拓扑关系矩阵$\boldsymbol{R}_{\mathrm{v}}$，其元素表示对应图像样本之间的相关性. 如图5所示，以第1个样本为例，通过排序算法可获得第1个样本图像最相似的k个图像，其对应的权重向量可表示为$\boldsymbol{r}^{1}\in\mathbb{R}^{k},k\ll N$. 第1个样本图像的表情标签分布表示为$\boldsymbol{D}^{1}=(d_{0}^{1},d_{1}^{1},\cdots,d_{C_f-1}^{1})$，其中，$d_{j}^{1}$为类别$j$的概率，且$\sum_{j=0}^{C_f-1}d_{j}^{1}=1$，其构建过程如式(18)所示：

$$\boldsymbol{D}^{1}=\frac{\sum_{i=1}^{k}(r_{i}^{1}\boldsymbol{P}_{i})}{\sum_{i=1}^{k}r_{i}^{1}} \tag{18}$$

其中，r_{i}^{1}表示与第1个样本最相似的第i个图像的相关性，$\boldsymbol{P}_{i}\in\mathbb{R}^{C_f}$表示和第1个图像最相似的第$i$个图像通过细粒度分类头得到的预测概率分布.

为了缓解单类别标签造成的标签歧义性对模型性能的影响，本文通过KL散度(Kullback-Leibler Divergence，KLD)约束每一批次图像的预测标签分布矩阵$\boldsymbol{P}$与近邻样本构造的标签分布矩阵$\boldsymbol{D}$的相似性，损失函数如式(19)所示：

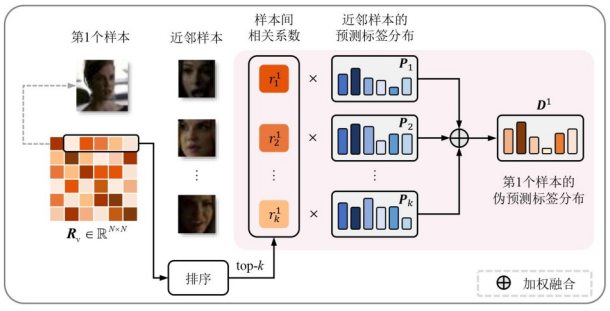


图5 基于近邻样本的伪预测标签分布生成模块示意

$$L_{\mathrm{KL}}=\mathrm{KL}(\boldsymbol{D}\|\boldsymbol{P})=-\frac{1}{NC_f}\sum_{i=0}^{N-1}\sum_{c=0}^{C_f-1}D_{c}^{i}\log(P_{c}^{i}) \tag{19}$$

3.5 损失函数

细粒度表情识别本质上是一个图像分类问题，因此本文采用交叉熵损失函数进行约束，如式(20)所示：

$$L_{f}=L_{\mathrm{CE}}(\hat{\boldsymbol{Y}},\boldsymbol{Y}) \tag{20}$$

其中,$\boldsymbol{Y}$是输入图像的细粒度真实标签,预测标签$\hat{\boldsymbol{Y}}$为预测分布$\boldsymbol{P}$中概率最大值对应的标签,即$\hat{\boldsymbol{Y}}=\text{argmax}(\boldsymbol{P})$,$L_{\text{CE}}$表示交叉熵损失函数. 本文的损失函数如式(21)所示:

$$L=L_f+L_{\text{KL}} \tag{21}$$

所提算法的整体流程如算法1所示:

算法1　基于关系感知和标签消歧的细粒度面部表情识别算法

输入:人脸图像$\boldsymbol{I}$,标签集合V,标签层级结构G,近邻样本数k,图像的真实标签$\boldsymbol{Y}$,最大迭代步N.

输出:图像的预测标签$\hat{\boldsymbol{Y}}$

方法:

FOR epoch=0 TO N−1 DO:
依据式(1)和式(2)分别获得图像特征$\boldsymbol{F}$和标签语义特征$\boldsymbol{S}$;
依据式(3)~(7)得到图像关系矩阵$\boldsymbol{R}_{\text{v}}$和图像拓扑表示$\tilde{\boldsymbol{F}}$;
依据式(10)~(12)得到标签特征$\tilde{\boldsymbol{S}}$;
遵循式(14)~(17)获得由标签特征引导后的图像特征$\hat{\boldsymbol{F}}$;
将$\hat{\boldsymbol{F}}$输入细粒度分类头得到图像的预测分布$\boldsymbol{P}$;
依据式(18),由图像的预测分布$\boldsymbol{P}$和图像关系矩阵$\boldsymbol{R}_{\text{v}}$得到此批次图像对应的标签分布$\boldsymbol{D}$;
计算图像的预测标签$\hat{\boldsymbol{Y}}=\text{argmax}(\boldsymbol{P})$;
依据式(19)计算$\boldsymbol{P}$和$\boldsymbol{D}$的KL散度损失;
依据式(20)计算$\hat{\boldsymbol{Y}}$和$\boldsymbol{Y}$的交叉熵损失;
通过优化式(21)的总损失函数更新网络参数;
END FOR

4　实验

4.1　数据集与评价指标

为了验证本文所提方法的有效性,本文在细粒度表情识别数据集FG-Emotions[25]上进行训练和测试. FG-Emotions包含来自于影视剧集中的10 371张图像,并包含6个粗粒度类别和33个细粒度类别. 本文采用FG-Emotions预处理后的关键帧数据,其中包括7 245张训练图像和3 126张测试图像. 此外,本文还在两个主流的7类粗粒度表情识别数据集AffectNet[37]和RAF-DB[38]上进行实验以进一步验证所提算法的有效性. 其中,AffectNet包括28 709张训练图像,3 589张验证图像和3 589张测试图像. RAF-DB的训练集包含12 271张图像,测试集包含3 068张图像. 本文利用平均准确率、各类别识别准确率,以及每秒处理帧数(Frames Per Second,FPS)作为评价指标,从而定量地评价RALD模型的识别结果.

4.2　实验设置

本文使用pytorch框架搭建模型,操作系统为Ubuntu 18.04.6 LTS,CPU为Hygon C86 7151 16-core Processor,GPU为NVIDIA RTX A4000,显存容量是16 GB.

在FG-Emotions数据集上,本文以ResNet18[39]为基础模型. 由于该数据集的粗细粒度标签命名存在重复,本文将每个粗粒度标签所属细粒度的特征进行平均,作为该粗粒度标签的语义特征. 实验的批处理大小为128. 式(8)中的λ取值为0.125. 式(9)中的σ设置为500. 本文将实验的初始学习率设置为0.01,使用SGD(Stochastic Gradient Descent,SGD)优化器迭代训练整个网络,每100轮对学习率进行衰减,衰减系数为5e^{-3},总共迭代300轮次.

4.3　实验结果

4.3.1　与现有方法比较结果

表1为RALD模型与主流方法在FG-Emotions数据集上的性能对比结果,RALD模型整体性能优于现有的主流方法. 平均准确率较EAC、DAN和POSTER分别高出约1.14%、5.00%和1.90%. 与通过图结构进行关系学习和推理的R3HO-Net相比,RALD模型提高了约1.51%的准确率,表明利用标签层级关系和图像特征优化有助于特征表征以及标签消歧. RALD在大多数类别上实现了更好的识别性能,验证了其优越性和鲁棒性. 此外,RALD的推理速度优于大多数方法,相较于DAN和POSTER分别快740 FPS和881 FPS,表明RALD模型在推理性能和推理速度上实现了较好的平衡.

另外,本文在两个基本表情识别数据集上对RALD模型的有效性进行验证,其结果如表2所示. 需要指出的是,由于基本表情数据集不存在标签层级关系,因此本文在RALD模型的基础上删除了层级结构编码部分,得到了算法退化版本RALD-L.

由表2可知,相比较于DACL、DACL+R3HO-Net、DAN、POSETR和REA-Net,本文所提RALD-L模型在RAF-DB的准确率分别提高了约2.80%、5.06%、0.88%、4.55%和1.77%. 在AffectNet上,RALD-L模型相比DACL、SCN和REA-Net分别提升了约0.04%、6.98%以及4.79%的准确率,验证了本文算法在缺少标签层级关系的条件下在基本表情数据集上的泛化性. 虽然RALD-L在AffectNet上相比于DAN和POSTER算法性能略有不足,但是本文设计的RALD-L算法旨在基于图像整体特征,深入探讨如何有效挖掘图像之间、标签之间以及图像与标签之间的相关性,并针对标签存在的歧义性问题展开研究. 需要特别指出的是,在基本表情数据集上,RALD-L即使不考虑层级结构编码,其性能依然优于R3HO-Net,侧面反映了本文利用注意力机制挖掘图像及标签关系和在标签消歧方法上的优越性. 由上述对比结果可知,本文所提算法的退化版本在基本表情数据集上相较于大多数方法具有一定的竞争性和优势.

4.3.2　消融实验

为了验证所提RALD算法中各个模块的有效性,本文采用未预训练的ResNet18作为基础模型(baseline),在细粒度面部表情识别数据集FG-Emotions上进行消融实验. 消融实验结果如表3所示.

方法(b)与基础模型相比,加入ITRR模块带来了约4.69%的提升,表明图像数据集的内在关系挖掘有助于

表1 RALD模型与主流方法在FG-Emotions数据集上的识别准确率和推理速度对比结果

积极或消极表情	粗粒度类	细粒度类	Resnet18[39]	MSAU-Net[25]	DACL[23]	SCN[40]	R18+R3HO[14]	EAC[19]	DAN[41]	POSTER[16]	RALD(Ours)
unpleasant	fear	anxiety	95.12	68.50	95.12	**97.56**	95.12	**97.56**	96.34	**97.56**	**97.56**
		worry	94.63	66.30	92.62	95.97	94.80	96.64	93.96	95.97	**97.99**
		panic	90.00	67.30	87.50	**97.50**	92.50	**97.50**	**97.50**	95.00	**97.50**
		terror	89.23	69.20	78.46	95.38	93.85	95.38	87.69	92.30	**96.92**
		nervous	**96.97**	67.80	**96.97**	93.94	93.94	**96.97**	93.94	**96.97**	**96.97**
	sadness	depression	92.80	70.20	95.20	96.80	95.20	95.20	95.20	98.40	**99.20**
		embarrass	91.52	65.90	95.76	98.23	98.23	98.23	96.82	98.58	**99.65**
		gloom	94.02	72.30	**98.91**	**98.91**	94.92	98.37	94.57	97.83	**98.91**
		guilt	94.92	68.20	**98.31**	**98.31**	**98.31**	94.92	91.53	96.61	**98.31**
		remorse	89.66	68.80	96.55	91.38	95.10	94.83	91.38	93.10	**98.27**
		sorrow	96.24	66.70	96.24	**98.50**	97.94	**98.50**	97.74	**98.50**	**98.50**
		suffering	93.55	78.00	90.32	93.55	**98.39**	98.39	88.71	90.32	**98.39**
	anger	anger	91.58	73.50	97.90	94.74	98.90	97.90	91.58	97.89	**98.95**
		contempt	85.39	72.90	92.14	**97.75**	94.38	95.51	92.14	**97.75**	**97.75**
		disgust	93.33	71.30	95.00	93.33	95.00	93.33	91.67	93.33	**96.67**
		frustrate	87.50	72.80	89.06	96.88	95.63	95.31	87.50	**98.44**	**98.44**
		hate	88.33	81.10	93.33	91.67	94.67	93.10	91.37	93.10	**96.55**
		irritate	91.89	76.40	92.66	97.68	**98.07**	97.30	93.05	97.68	**98.07**
		loathing	85.85	77.90	91.51	92.45	**96.22**	92.45	86.79	91.51	**96.22**
		torment	91.80	70.10	90.16	93.44	93.44	95.08	90.16	95.08	**97.72**
		wrath	92.86	70.20	83.33	88.10	96.00	**97.62**	85.71	85.71	95.08
pleasant	surprise	amaze	89.45	85.80	89.87	91.98	93.79	91.98	92.40	93.67	**96.20**
		astonish	90.20	74.60	90.20	90.19	92.16	**96.08**	88.24	92.16	**96.08**
		surprise	86.87	77.40	83.84	89.90	89.90	**90.91**	79.80	**90.91**	**90.91**
	happiness	cheerful	94.34	69.80	96.22	**99.06**	98.17	**99.06**	97.17	**99.06**	98.11
		enthrall	89.47	82.70	94.74	**98.25**	**98.25**	96.49	94.74	**98.25**	**98.25**
		optimism	95.54	75.40	92.99	**99.36**	97.45	98.73	96.18	**99.36**	98.73
		pleasure	93.21	73.80	95.68	98.14	97.83	98.14	96.91	97.53	**98.77**
		pride	95.24	76.80	85.71	95.24	90.48	90.48	**95.24**	**95.24**	**95.24**
		relief	85.71	75.70	**95.24**	90.48	90.48	**95.24**	**95.24**	90.48	**95.24**
		thrill	97.50	80.30	95.00	**97.50**	**97.50**	95.00	95.00	**97.50**	**97.50**
	love	affection	95.00	83.30	95.00	92.50	95.00	**97.50**	92.50	**97.50**	**97.50**
		lust	88.46	82.10	92.31	92.31	80.77	**96.15**	88.46	**96.15**	**96.15**
平均准确率/%		Avg.	91.91	73.73	92.54	95.06	95.83	95.93	92.34	95.44	**97.34**
推理速度/FPS		Speed	1 942	—	1 902	1 873	—	318	551	410	1 291

表2 RALD-L模型在粗粒度表情识别数据集AffectNet和RAF-DB上与主流方法的识别准确率对比 单位:%

粗粒度表情数据集	IL-CNN[42]	MSAU-Net[25]	DACL[23]	DACL+R3HO-Net[14]	SCN[40]	Ada-CM[43]	DAN[41]	POSTER[16]	REA-Net[44]	RALD-L (Ours)
AffectNet	56.60	—	65.20	58.26	58.45	52.97	65.69	**67.31**	60.45	65.24
RAF-DB	82.30	75.80	87.78	85.52	82.45	84.13	89.70	86.03	88.81	**90.58**

获得更具辨别力的图像特征表示. 方法(*d*)通过LGIFO模块引入标签特征后,相较于基础模型准确率提升了4.02%,这表明利用标签自身语义特征与图像之间的关联性有助于面部表情识别准确率的进一步提升. 方法(*e*)和方法(*f*)分别挖掘了图像关系和标签关系,都可以在方法(*d*)的基础上实现提升,验证了本文两种关系感知模块的有效性. 最后,方法(*g*)通过引入基于近邻样本的标签分布学习模块(NNLDL),识别准确率得到进一步提升,验证了标签分布学习能有效缓解标签歧义性对模型性能的影响. 整体模型相比于基础模型在识别准确率上提高了约5.43%,验证了所提各模块的有效性.

表3 FG-Emotions数据集上的消融实验

方法变体		ITRR	HLRR	LGIFO	NNLDL	准确率/%
without label guided	(*a*)					91.91
	(*b*)	√				96.60
	(*c*)	√			√	96.45
with label guided	(*d*)			√		95.93
	(*e*)	√		√		96.80
	(*f*)		√	√		96.06
	(*g*)	√	√	√		97.25
Ours		√	√	√	√	**97.34**

针对FG-Emotions数据集,在基于近邻样本的标签分布学习模块中使用不同近邻样本数量k的消融实验如图

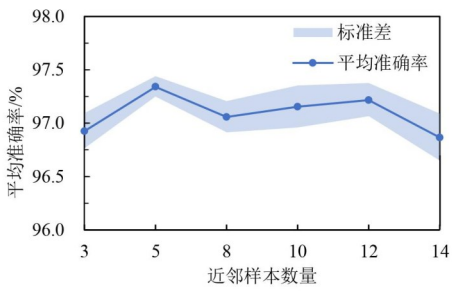


图6　不同近邻样本数量的影响

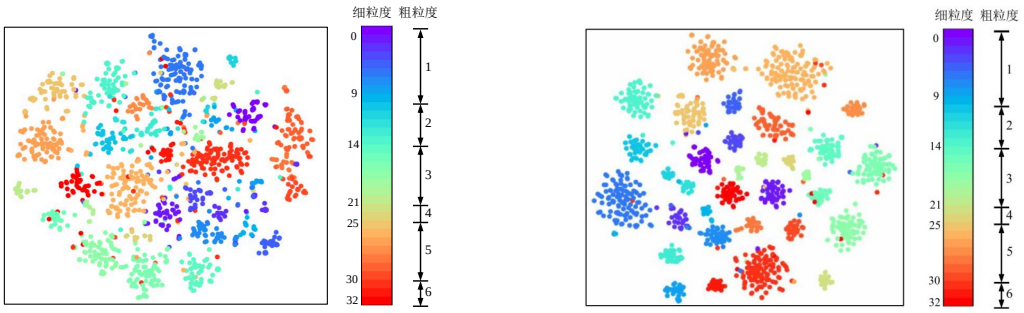
6所示. 当近邻样本k=5时,识别准确率最高. 随着k的继续增大,准确率下降了约0.28%. 这表明,k过大可能会导致冗余信息,继而降低模型性能. 而k较小时,则无法充分利用近邻样本的信息,导致所学伪预测标签分布可靠性不足,难以为预测标签分布提供有效指导.

4.3.3　可视化分析

为验证RALD模型的有效性,本文进行了定性的可视化分析. 图7展示了基线模型与RALD模型在FG-Emotions数据集上细粒度表情识别的对比结果. 结果表明,细粒度表情识别相较于基本表情识别更具挑战性. 同时,与基线模型相比,RALD模型识别结果更加准确. 如图7右侧的第二个失败案例所示,即便在有面部遮挡情况下,RALD的识别结果"嫌恶"仍与真实标签"憎恨"属于同一粗粒度类,而基线模型的结果"忧郁"则属于粗粒度类"悲伤",体现了本文通过挖掘粗细粒度标签之间的层级关系对面部表情识别的有效性. 图8展示了当近邻数量为5时,基于近邻样本的标签分布学习模块中最相似近邻样本的图像以及相似权重值. 通过图像预测分布的加权融合,可以有效地利用数据集特征空间的近邻相似性信息,从而更准确地构造样本的伪预测标签分布. 相比于单一标签,这种表情类别分布可以为模型的学习提供更丰富的监督信息,进而缓解标签歧义性对模型性能的影响,提高面部表情识别的准确性和鲁棒性.

图7　RALD模型与基线模型的识别结果可视化对比

图8　基于近邻样本的标签分布学习模块中top-5样本及权重值示例



(*a*) 基线模型结果 (*b*) RALD模型结果

图9 t-SNE可视化结果对比

此外,本文还使用t-SNE[45]进行特征空间可视化.本文从FG-Emotions测试集中随机选取2 500张图像,分别使用基线模型和RALD模型进行特征提取.图9为二者的特征分布进行t-SNE可视化后的对比结果.其中,33类细粒度标签由颜色条表示,而6个黑色箭头指示了粗粒度标签涵盖的细粒度标签范围.由图9可知,相比于基线模型,RALD模型对每一类别的分类结果形成了更紧凑的簇,反映了所学特征具有更强的可辨别性.需要特别指出的是,从粗粒度分类的角度,RALD模型结果中属于同一粗粒度的不同细粒度类别在特征空间中距离更近.这表明在没有训练粗粒度分类器的情况下,RALD模型不仅可以有效分离细粒度类别,还可以使其满足粗粒度类别的约束关系,体现了RALD模型挖掘粗细粒度关系的有效性和优越性.

5 结语

本文提出了一种基于关系感知和标签消歧的细粒度面部表情识别算法RALD,其利用层级感知的图像特征增强网络,挖掘图像结构关系和标签层级关系,从而获得图像拓扑关系表示和层级感知的标签关系表示,并利用标签信息引导图像特征进行优化,以此提升模型的识别性能.此外,本文提出基于近邻样本的标签分布学习模块,基于特征空间的近邻图像信息构造样本的标签分布,以此缓解标签歧义问题.在细粒度表情识别数据集FG-Emotions和主流粗粒度表情识别数据集AffectNet和RAF-DB上,大量的消融和对比实验结果验证了我们所提出的RALD算法的有效性和优越性.

参考文献

[1] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664-676.

[2] ZHANG H, KOH J Y, BALDRIDGE J, et al. Cross-modal contrastive learning for text-to-image generation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 833-842.

[3] LIU T T, WANG J X, YANG B, et al. Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom[J]. Infrared Physics and Technology, 2021, 112: 103594.

[4] AGRAWAL A, LU J S, ANTOL S, et al. VQA: Visual question answering[J]. International Journal of Computer Vision, 2017, 123(1): 4-31.

[5] EKMAN P, FRIESEN W V. Constants across cultures in the face and emotion[J]. Journal of Personality and Social Psychology, 1971, 17(2):124-129.

[6] LEE J, KIM S, KIM S, et al. Context-aware emotion recognition networks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 10143-10152.

[7] 姚乃明, 郭清沛, 乔逢春,等. 基于生成式对抗网络的鲁棒人脸表情识别[J]. 自动化学报, 2018, 44(5): 865-877.
YAO N M, GUO Q P, QIAO F C, et al. Robust facial expression recognition with generative adversarial networks[J]. Acta Automatica Sinica, 2018, 44(5): 865-877. (in Chinese)

[8] LIU Z Y, YUAN X Y, LI Y T, et al. PRA-Net: Part-and-relation attention network for depression recognition from facial expression[J]. Computers in Biology and Medicine, 2023, 157: 106589.

[9] 孙晓, 潘汀. 基于兴趣区域深度神经网络的静态面部表情识别[J]. 电子学报, 2017, 45(5): 1189-1197.
SUN X, PAN T. Static facial expression recognition system using ROI deep neural networks[J]. Acta Electronica Sinica, 2017, 45(5): 1189-1197. (in Chinese)

[10] 张瑞, 蒋晨之, 苏剑波. 基于稀疏特征挑选和概率线性判别分析的表情识别研究[J]. 电子学报, 2018, 46(7): 1710-1718.
ZHANG R, JIANG C Z, SU J B. Expression recognition based on sparse selection and PLDA[J]. Acta Electronica Sinica, 2018, 46(7): 1710-1718. (in Chinese)

[11] ZHANG W, JI X P, CHEN K Y, et al. Learning a facial

expression embedding disentangled from identity[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 6759-6768.

[12] 章思远, 肖世明, 张蓬, 等. 图像生成和深度度量学习的身份感知面部表情识别方法[J]. 计算机辅助设计与图形学学报, 2021, 33(5): 724-732.
ZHANG S Y, XIAO S M, ZHANG P, et al. Identity-aware facial expression recognition method based on synthesized images and deep metric learning[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(5): 724-732. (in Chinese)

[13] ZENG J B, SHAN S G, CHEN X L. Facial expression recognition with inconsistently annotated datasets[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 227-243.

[14] ZHU Y C, WEI L L, LANG C Y, et al. Fine-grained facial expression recognition via relational reasoning and hierarchical relation optimization[J]. Pattern Recognition Letters, 2022, 164: 67-73.

[15] CAI J, MENG Z B, KHAN A S, et al. Probabilistic attribute tree structured convolutional neural networks for facial expression recognition in the wild[J]. IEEE Transactions on Affective Computing, 2023, 14(3): 1927-1941.

[16] ZHENG C, MENDIETA M, CHEN C. POSTER: A pyramid cross-fusion Transformer network for facial expression recognition[C]//2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE, 2023: 3146-3155.

[17] QI Y F, ZHOU C Y, CHEN Y X. NA-Resnet: Neighbor block and optimized attention module for global-local feature extraction in facial expression recognition[J]. Multimedia Tools and Applications, 2023, 82(11): 16375-16393.

[18] CHEN D L, WEN G H, LI H H, et al. Multi-relations aware network for in-the-wild facial expression recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 3848-3859.

[19] ZHANG Y H, WANG C R, LING X, et al. Learn from all: erasing attention consistency for noisy label facial expression recognition[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 418-434.

[20] XUE F L, WANG Q C, GUO G D. TransFER: Learning relation-aware facial expression representations with Transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 3601-3610.

[21] 胡敏, 江河, 王晓华, 等. 基于几何和纹理特征的表情层级分类方法[J]. 电子学报, 2017, 45(1): 164-172.
HU M, JIANG H, WANG X H, et al. A hierarchical classification method of expressions based on geometric and texture features[J]. Acta Electronica Sinica, 2017, 45(1): 164-172. (in Chinese)

[22] 廖海斌, 徐斌. 基于性别和年龄因子分析的鲁棒性人脸表情识别[J]. 计算机研究与发展, 2021, 58(3): 528-538.
LIAO H B, XU B. Robust face expression recognition based on gender and age factor analysis[J]. Journal of Computer Research and Development, 2021, 58(3): 528-538. (in Chinese)

[23] LIU X F, KUMAR B V K V, YOU J, et al. Adaptive deep metric learning for identity-aware facial expression recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2017: 20-29.

[24] FARZANEH A H, QI X J. Facial expression recognition in the wild via deep attentive center loss[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2021: 2402-2411.

[25] LIANG L Q, LANG C Y, LI Y D, et al. Fine-grained facial expression recognition in the wild[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 482-494.

[26] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL] (2021-06-03)[2024-04-22]. https://arxiv.org/abs/2010.11929.

[27] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 213-229.

[28] ZHENG S X, LU J C, ZHAO H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 6881-6890.

[29] XUE F L, WANG Q C, TAN Z C, et al. Vision Transformer with attentive pooling for robust facial expression recognition [J]. IEEE Transactions on Affective Computing, 2023, 14(4): 3244-3256.

[30] GENG X. Label distribution learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(7): 1734-1748.

[31] GAO B B, XING C, XIE C W, et al. Deep label distribution learning with label ambiguity[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2825-2838.

[32] LE N, NGUYEN K, TRAN Q, et al. Uncertainty-aware label distribution learning for facial expression recognition[C]// 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2023: 6088-6097.

[33] CHEN S K, WANG J F, CHEN Y D, et al. Label distribution learning on auxiliary label space graphs for facial expression recognition[C]//2020 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 13984-13993.
[34] ZHAO Z Q, LIU Q S, ZHOU F. Robust lightweight facial expression recognition network with label distribution training[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3510-3519.
[35] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 2013.
[36] YING C, CAI T, LUO S J, et al. Do Transformers really perform badly for graph representation? [C]//Neural Information Processing Systems, 2021, 34: 28877-28888.
[37] MOLLAHOSSEINI A, HASANI B, MAHOOR M H. AffectNet: A database for facial expression, valence, and arousal computing in the wild [J]. IEEE Transactions on Affective Computing, 2017, 10: 18-31.
[38] LI S, DENG W H, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2852-2861.
[39] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.
[40] WANG K, PENG X J, YANG J F, et al. Suppressing uncertainties for large-scale facial expression recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 6897-6906.
[41] WEN Z Y, LIN W Z, WANG T, et al. Distract your attention: Multi-head cross attention network for facial expression recognition[J]. Biomimetics, 2023, 8(2): 199.
[42] CAI J, MENG Z B, KHAN A S, et al. Island loss for learning discriminative features in facial expression recognition[C]// 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Piscataway: IEEE, 2018: 302-309.
[43] LI H Y, WANG N N, YANG X, et al. Towards semi-supervised deep facial expression recognition with an adaptive confidence margin[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4166-4175.
[44] 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 G G, ZHANG F, WANG H, et al. Facial expression recognition based on region enhanced attention network[J]. Journal of Computer-Aided Design & Computer Graphics, 2024, 36(1): 152-160. (in Chinese)
[45] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2625.

作者简介

刘雅芝 女,硕士研究生. 主要研究方向为计算机视觉和人脸表情识别.
E-mail: 22120395@bjtu.edu.cn

许喆铭 女,博士研究生. 主要研究方向为车辆重识别和多视图学习.
E-mail: 21112016@bjtu.edu.cn

郎丛妍 女,博士. 教授,博士生导师,CCF会员. 主要研究方向为计算机视觉和多媒体内容分析.
E-mail: cylang@bjtu.edu.cn

王 涛 男,博士. 教授. 主要研究方向为计算机视觉和机器学习.
E-mail: twang@bjtu.edu.cn

李浥东 男,博士. 教授,博士生导师. 主要研究方向为大数据分析、隐私保护、信息安全和数据挖掘.
E-mail: ydli@bjtu.edu.cn