Computer Science and Application 计算机科学与应用, 2026, 16(7), 154-168
Published Online July 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.167249

Hans汉斯

CEENet：结合对比学习与视觉嵌入的情绪识别方法

夏旭辉，田春岐*

同济大学计算机科学与技术学院，上海

收稿日期：2026年6月19日；录用日期：2026年7月17日；发布日期：2026年7月24日

摘　要

针对当前上下文情绪识别研究中抽象情绪语义学习不足、上下文线索融合方式单一以及面部表情特征分离质量不佳等困难，文章提出了一种结合对比学习与视觉嵌入增强的情绪识别网络CEENet (Context Embedding Emotion Network)。该网络通过并行提取背景、身体、面部以及基于视觉－语言模型生成的视觉嵌入共四路特征作为情绪识别的上下文，进一步加深了对高级情绪语义的挖掘。为实现多来源上下文特征的高效融合，文章设计了多流特征注意力模块，利用自适应注意力机制动态调整各上下文信息在情绪识别中的权重，强化情绪相关线索并抑制噪声。此外，针对面部表情特征的分离，文章提出了一种多尺寸特征增强的面部表情嵌入网络，并引入一种创新的三层级数据增强三元表情对比学习方案，显著提升了表情嵌入的判别力。实验结果表明，CEENet在Emotic和CAER-S等上下文情绪识别基准测试上分别达到了34.15%的平均精确率均值和88.25%的分类准确率，同时本研究中的面部表情嵌入网络也在FEC基准测试中取得了86.1%的准确率，均优于同类型主流方法。最后，消融实验进一步验证了主干特征提取结构、特征融合结构及多层级数据增强策略在提升模型性能方面的有效性。

关键词

情绪识别，表情嵌入，对比学习，数据增强，多尺寸特征增强

CEENet: Emotion Recognition with Contrastive Learning and Visual Embeddings

Xuhui Xia, Chunqi Tian*

Department of Computer Science and Technology, Tongji University, Shanghai

Received: June 19, 2026; accepted: July 17, 2026; published: July 24, 2026

*通讯作者。

文章引用: 夏旭辉, 田春岐. CEENet: 结合对比学习与视觉嵌入的情绪识别方法[J]. 计算机科学与应用, 2026, 16(7): 154-168. DOI: 10.12677/csa.2026.167249

Abstract

To address the challenges in current context-aware emotion recognition research, including inadequate learning of abstract emotional semantics, limited strategies for fusing contextual cues, and unsatisfactory disentanglement of facial expression features, this paper proposes CEENet (Context Embedding Emotion Network), an emotion recognition network that integrates contrastive learning with visual embedding enhancement. CEENet extracts four parallel streams of features as contextual cues for emotion recognition, namely background, body, face, and visual embeddings generated by a vision-language model, thereby enhancing the modeling of high-level emotional semantics. To achieve efficient fusion of multi-source contextual features, a Multi-Stream Feature Attention module is designed to dynamically adjust the weights of different contextual information through an adaptive attention mechanism, thereby strengthening emotion-relevant cues while suppressing noise. In addition, to address the disentanglement of facial expression features, this paper proposes a facial expression embedding network with multi-scale feature enhancement and introduces an innovative three-level data-augmentation-based triplet expression contrastive learning scheme. The proposed method significantly improves the discriminability of expression embeddings. Experimental results show that CEENet achieves a mean average precision of 34.15% and a classification accuracy of 88.25% on context-aware emotion recognition benchmarks such as Emotic and CAER-S, respectively. Meanwhile, the facial expression embedding network developed in this study attains an accuracy of 86.1% on the FEC benchmark, outperforming existing mainstream methods. Finally, ablation studies further verify the effectiveness of the backbone network design, the fusion network design, and the multi-level contrastive learning strategy in improving emotion recognition performance.

Keywords

Emotion Recognition, Expression Embedding, Contrastive Learning, Data Augmentation, Multi-Scale Feature Enhancement

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着人机交互和人工智能技术的飞速发展，人脸分析已成为计算机视觉领域的核心研究方向之一，其在身份认证、情感计算、智能监控等场景中扮演着至关重要的角色。其中人脸表情识别专注于解读面部肌肉运动所传达的情感状态，简称 FER。自上世纪 70 年代 Ekman 与 Friesen [1]开创性地验证了情绪表达的相通性，后续研究提出六种跨文化的基本情绪[2]以来，FER 大多围绕这些离散的情感类别展开，并逐渐成为理解人类行为与意图的关键技术。

近年来，以卷积神经网络为代表的深度学习技术凭借其强大的特征自动学习能力，已成为 FER 任务的主流范式，过去情绪识别的方法主要聚焦于面部图像的分析[3]，但心理学研究表明，人类情绪的表达具有多种来源——肢体姿态(如低头耸肩)、场景上下文(如聚会场景)等非面部线索同样承载着关键情绪信息[4]。为此，EMOTIC [5]工作率先提出了“上下文情绪识别”任务，要求系统同时建模面部和人体，背景等其他特征，提取其中与情绪相关的信息线索，综合进行情绪识别。另外，针对面部情绪特征的研究表明，人脸的视觉特征本质上是一种高度耦合的信号，其中与情感相关的特征常常与个体的生理结构特

征等其他特征交织在一起，为了学习到纯粹的、仅与表情相关的表征，研究人员致力于使用如图 1 所示的三元对比学习的方式[6] [7]，将高维的人脸图像信息压缩到一个紧凑且具有判别力的低维特征空间中，学习到表情嵌入向量，进而用于下游任务例如情绪识别、表情检索等。

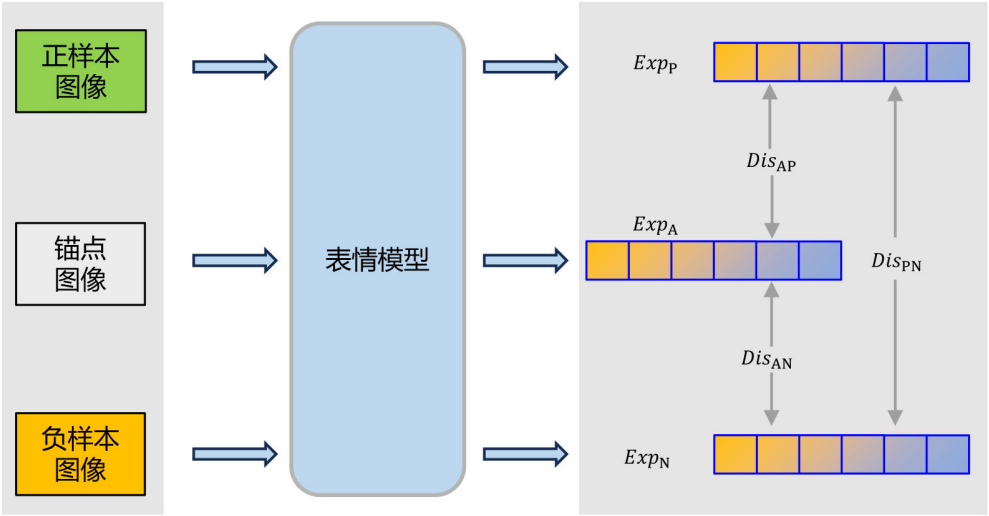

Figure 1. Process of the expression triplet's contrastive learning
图 1. 表情三元组对比学习流程

然而，现有的 FER 方法在三个方面存在局限：(1) 对上下文语义的理解停留在低级视觉特征层面，未能充分利用图像所体现的如被识别人所进行的活动等抽象语义。(2) 模型难以判断不同上下文特征与情绪识别任务的相关度，情绪无关特征往往容易影响最终结果。(3) 受训练方法，网络结构等因素影响，面部情绪表征在三元对比学习中的分离质量不高，进而降低了情绪识别的效果。

针对上述问题，本文提出上下文嵌入增强的情绪识别网络 CEENet (Context Embedding Emotion Network)，CEENet 同时提取图像的背景，身体，面部以及视觉嵌入四项上下文特征，通过自适应注意力模块将多种不同特征进行加权融合，同时在其中设计了新的面部情绪特征提取子网络。最终在上下文情绪识别基准测试以及面部表情对比基准测试上获得了优于主流方法的指标，在上下文情绪识别领域处于领先水平。主要工作包括：

1) 引入视觉嵌入上下文：在传统上下文情绪识别方法所用到的背景 - 身体 - 面部三类上下文之外，使用视觉 - 语言模型提取图像的视觉嵌入作为新的上下文并参与到后续的情绪识别中，视觉嵌入中的场景语义信息可以有效补充环境下的情绪线索，进而提高了情绪识别的总体性能。

2) 提出多路特征注意力模块 MFA (Multi-Stream Feature Attention)，根据图像信息计算各类上下文为情绪分类结果的贡献权重，使得网络加强了对情绪相关特征的重视程度的同时弱化了情绪无关特征，进而提高了情绪识别能力。

3) 使用一种三层级数据增强的三元对比学习方法进行预训练，进而得到更高性能的面部表情表征提取子网络 MEEN (Multi-Scale Expression Embedding Network)。该方案不仅通过传统的图像变换丰富了视觉多样性，还利用人脸图像与自身的不同视图的表情不变性以及样本间的偏序关系挖掘了新的监督信号，显著扩充了训练数据的等效规模与复杂度，从而引导模型学习到更加纯粹的表情嵌入向量。

本文的其余部分安排如下：第二节介绍本研究的先前基础工作，第三节论述本研究所提出的具体方法，第四节介绍本研究所进行的实验方案与结果，第五节就实验结果进行总结并阐述本文结论。

2. 相关工作

2.1. 上下文情绪识别

情绪识别的早期工作集中于面部表情识别，Matsumoto 等[2]提出的六种基本情感分类奠定了FER2013 [8]等基准数据集的理论基础。然而，单纯依赖面部特征难以解释真实场景中的复杂情感、遮挡表情等，这促使研究者引入身体姿态与场景上下文[9] [10]作为补充特征。2017 年 Kosti 等[5]构建的EMOTIC 数据集首次将场景上下文纳入情感识别任务，并验证了上下文特征对 valence-arousal 维度[11]预测的重要性。在这之后，多流网络架构成为上下文情绪识别的主流解决方案，Ninh 等[12]在使用人脸，身体与背景三类特征进行情绪识别，Mittal 等[13]除了使用前述特征之外，还通过图神经网络建模图片内多人的相互关系，后来这一方法也被推广至多模态情绪识别[14]中。然而，这些方法存在两方面缺陷：其一，上下文编码器局限于传统特征提取模型，缺乏对抽象语义(如“孤独”、“压抑”)的推理能力；其二，在多路特征融合时往往在拼接后通过全连接层进行直接分类，难以适应不同场景下的模态重要性变化。

2.2. 视觉－语言模型

以 CLIP [15]为代表的视觉－语言预训练模型，通过海量图文对的对比学习实现了跨模态语义空间的统一表征，为多模态学习提供了新的技术路径。相比传统视觉模型，CLIP 通过自然语言监督信号建立视觉概念与语义描述的联系，使模型能够更容易地理解“温馨”、“恐怖”等抽象属性。在这之后，图像生成，视频算法，音频对齐，3D 图像以及目标检测等方面涌现了一批跨越性的工作[16]-[19]，而另外一部分工作则在 CLIP 的基础上扩展了其能力，如 BLIP 与 BLIP-2 [20] [21]通过提出新的预训练方法减少了灾难性遗忘现象。CLIP-MoE [22]通过分阶段微调来得到具有互补信息的 FFN，并以此构建 MoE，获得了显著减少的训练与推理成本。Qwen2.5VL 系列模型[23]通过改进位置编码以及在视觉层增加特征融合连接等方法，达到了开源 MLLM 中的领先地位。而在情绪识别领域，人们也在探索跨模态的图文模型在情绪识别中的应用潜力，EmotionCLIP [24]则尝试将 CLIP 引入上下文情绪识别任务并且取得了赛道领先的成绩。

本研究中同样用到了 CLIP 图像编码器以增强情绪识别任务性能，但与 EmotionCLIP 的工作有较大差别。EmotionCLIP 主要研究改进 CLIP 的预训练，从而得到更好的视觉情感表示，属于情绪识别的上游任务，本文则主要研究如何利用 CLIP 的迁移学习能力，直接为情绪识别下游任务提供额外情感信息，进而提高模型的上下文情绪识别能力。同时，本文重新设计了上下文情绪识别网络结构，且进一步针对面部上下文的表情提取进行了基于对比学习的优化，在更细粒度的层面进行了创新。

2.3. 表情嵌入学习

表情表征，或者表情嵌入旨在将高维的人脸图像中的表情信息压缩到一个紧凑且具有判别力的低维特征空间中。为了适应表情特征的非离散性，研究者们开始探索如何直接学习表情之间的视觉相似度从而学习度量空间。其中，Vemulapalli 等人提出的 FECNet [6]首次构建了一个大规模的人脸表情对比数据集 FEC，其中包含了数十万个人工标注的表情图像三元组，每个三元组中包括锚点图像，正样本图像(相似表情图像)以及负样本图像(不相似表情图像)。通过在这个数据集上进行如图 1 所示的三元组对比学习，FECNet 成功地学习到了一个 16 维的紧凑表情嵌入。该嵌入在表情检索、表情分类等任务上表现出色，证明了通过对比学习直接建模表情嵌入的可行性。为了显式地解决表情与身份特征的纠缠问题，Zhang 等人提出了 DLN [7]，基于一个人的表情特征可以被看作是其面部总特征相对于其“中性”身份特征的一种“偏离”的假设，DLN 在预训练中加入视觉特征与身份特征的双分支网络与减法操作进而获取到了优于 FECNet 的表情嵌入质量。然而，现有的表情嵌入对比学习方法仍在数据增强策略、网络结构上存在一些

较多的进步空间，表情嵌入的质量有待提升。

3. 方法

本章将介绍本文的核心工作 CEENet，首先将从总体层面阐述 CEENet 网络的总体设计，具体描述其中所使用的网络结构的组成与改进，并介绍多流特征注意力模块的具体实现方式，最后对新型对比学习方法以及作为其训练成果的面部情绪网络进行详细介绍。

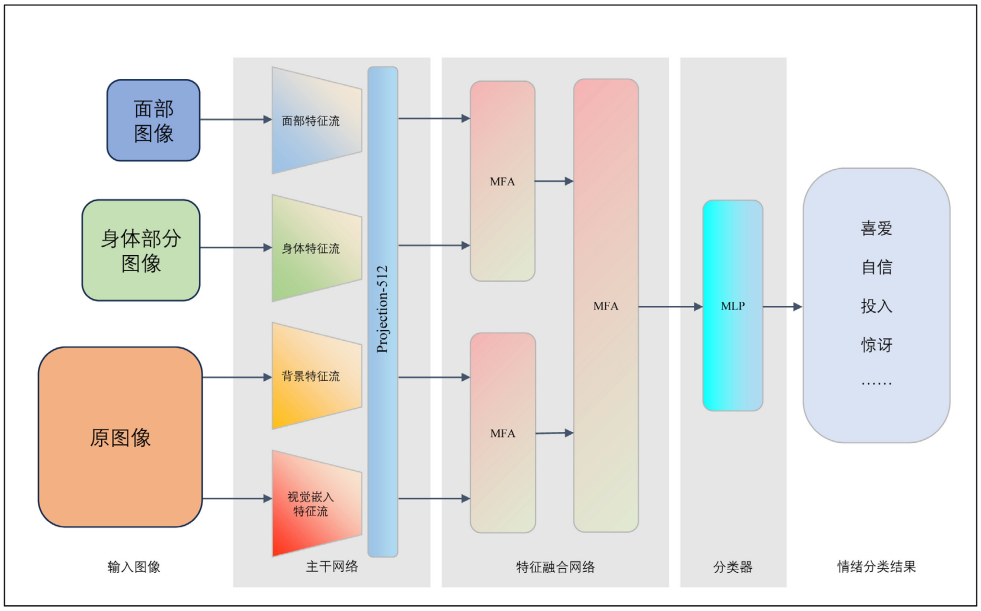

Figure 2. Schematic diagram of the CEENet structure
图 2. CEENet 结构示意图

3.1. CEENet 总体结构

本研究提出如图 2 所示的多流上下文情绪识别网络。网络的主要工作阶段包括图像输入，特征提取，特征融合，情绪分类等，下面将从主干网络与特征融合网络两部分来介绍 CEENet 的整体架构。

主干网络：由于图像各上下文中情绪线索有着不同表现形式，如面部主要由五官与肌肉动作表现情绪，身体主要由姿态表现情绪，整体场景则由被检测人的活动、背景氛围等表现情绪，而从这些不同的表现形式中提取情绪线索相当于不同的模式识别任务，因此本研究为不同上下文的特征提取针对性地设置了异构的编码器，以确保情绪线索的有效挖掘。主干网络包含视觉嵌入编码器、背景编码器、身体特征提取器以及面部特征提取器，输入图像经区域分割后送入各编码器分别处理，最后均通过各自的一层 MLP 投影到 512 维度的特征向量，编码器具体配置如下：

1) 全局图像经过尺寸调整后分别输入背景提取网络以及视觉嵌入提取网络，本方法中，背景特征编码器使用 ResNet50 [25]，视觉嵌入编码器使用 CLIP 系列模型中 ViT-B/32 规格的图像编码器。

2) 身体区域经尺寸调整后输入身体特征提取网络，本方法中使用 Tiny Swin-Transformer [26]作为编码器。

3) 面部区域经尺寸调整与灰度处理后，输入面部特征提取网络，本方法针对面部上下文子网络设计了多尺寸表情嵌入网络 MEEN (Multi-Scale Expression Embedding Network)，并经过改进后的三元对比学习预训练作为面部特征编码器，该网络的训练细节与设计将会在 2.3 节与 2.4 节中详细介绍。

特征融合网络：主干网络提取图像各类特征之后，需要将其融合后才能送入分类层。在本方法中我们选择先通过 MFA 模块将背景上下文与视觉嵌入上下文融合，同时将身体上下文与面部信息上下文，从四类特征变换为两类特征(环境上下文与主体上下文)之后，再使用一层 MFA 将这两类特征进行融合，最终得到待分类张量。待分类张量进入分类器后计算得到该图像处于每类情感的概率。以上过程的总体数学表达式为：

$$P=\sigma\left(\mathcal{F}_{\mathrm{MLP}}\left(\mathcal{F}_{\mathrm{fusion}}\left(\boldsymbol{f}_{\mathrm{back}},\boldsymbol{f}_{\mathrm{caption}},\boldsymbol{f}_{\mathrm{body}},\boldsymbol{f}_{\mathrm{face}}\right)\right)\right) \tag{1}$$

其中表示 Sigmoid 函数，为自适应融合网络，将融合后的特征记为，则展开计算过程后得

$$\boldsymbol{f}_{\mathrm{fusion}}=\mathcal{F}_{\mathrm{MFA}}\left(\mathcal{F}_{\mathrm{MFA}}\left(\boldsymbol{f}_{\mathrm{back}},\boldsymbol{f}_{\mathrm{caption}}\right),\mathcal{F}_{\mathrm{MFA}}\left(\boldsymbol{f}_{\mathrm{body}},\boldsymbol{f}_{\mathrm{face}}\right)\right) \tag{2}$$

3.2. 多流特征注意力模块

本研究为了提高 CEENet 分辨上下文信息对情绪识别的有效性的能力，强化情绪相关上下文，弱化情绪无关上下文，提出了自适应多流特征注意力模块 MFA (Multi-Stream Feature Attention)，用于多类不同上下文的加权融合。如图 3，从多个编码器分别获取形状不一的一维张量特征后，我们选取它们的长度的一个公因子(Common Divisor)

$$L=\mathcal{F}_{\mathrm{CD}}\left(\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3,\cdots\right) \tag{3}$$

将所有特征一维张量变形为第一维度为 L 的二维张量并在第二维拼接，得到一个 $L\times C$ 的张量。本研究所使用的 L 值取 128，L 值的选用主要来自参数搜索。

$$\boldsymbol{f}_{\mathrm{L\times C}}=\mathcal{F}_{\mathrm{Reshape}}\left(\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3,\cdots|\boldsymbol{L}\right) \tag{4}$$

考虑到每个上下文中所蕴含的情绪线索的连续性，每一个编码器得到的一维特征内都有情绪无关的噪声，也存在情绪相关的重要信息，因此不应简单地对四类上下文进行加权，而应该更加详细地对每一特征的每一部分进行重要性加权。因此本研究通过一个公因子来将各个上下文对应的情绪特征划分成通道数量为 C 的一维多通道张量，并尝试计算它们所属通道的重要性。

之后，将这个张量沿第一维进行平均池化操作，得到一个 $1\times C$ 的张量，再经过多层感知机以及 softmax 函数，得到 $\boldsymbol{f}_{L\times C}$ 中每个通道所对应的权重即注意力：

$$\boldsymbol{Attn}_{\mathrm{1\times C}}=soft\max\left(F_{MLP}\left(F_{Avgpool}\left(\boldsymbol{f}_{L\times C}\right)\right)\right) \tag{5}$$

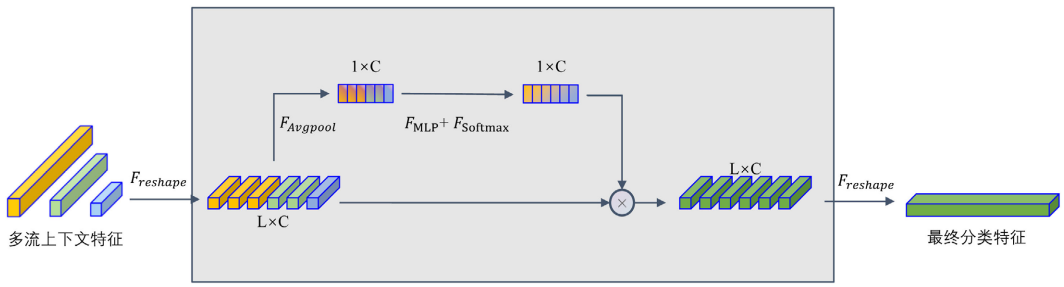

Figure 3. Schematic diagram of multi-stream feature attention module workflow
图 3. 多流特征注意力模块工作流程示意图

最后，将原始特征张量与注意力张量逐通道相乘，转换为待分类张量，完成 MFA 模块的计算：

$$\boldsymbol{f}_{\text{cls}} = \boldsymbol{f}_{L\times C} \odot \boldsymbol{Attn}_{1\times C} \tag{6}$$

在上述一系列计算过程中，MFA 模块给了神经网络一个为输入特征增加权重的渠道，在反向传播的过程中，这种自注意力机制可以逐渐识别到对分类结果贡献较多的特征并赋予其更高的重要程度，在一定程度上提高了网络的整体性能。

3.3. 三元表情对比学习预训练

如引言图 1 所示，本文基于三元对比学习对 3.4 节中介绍的面部情绪特征提取网络 MEEN 进行预训练，在介绍本研究对该对比学习过程进行改进前，此处先对三元表情对比学习的样本定义进行详细解释。

以本节所使用的面部表情对比数据集 FEC (Facial Expression Comparison)为例，其标注不提供情感类别，而是向标注者展示三张人脸图像，要求其从三种可能的图像对组合中选出表情最相似的一对。三元组代表锚点图像 a 与正样本图像 p 构成表情最相似的一对，同时也代表“p 图像比 n 图像更像 a 图像”，针对这三张图像中提取的表情嵌入向量，使用余弦相似度可以表示为：

$$S_{\cos}\left(\boldsymbol{f}_a, \boldsymbol{f}_p\right) > S_{\cos}\left(\boldsymbol{f}_a, \boldsymbol{f}_n\right) \tag{7}$$

在预训练过程中，通过三元对比损失来保持这个不等式约束，进而能够学习到与表情相似度密切绑定的度量空间，在这个度量空间中，相似的表情图像所提取的嵌入向量的相似度更高，反之则更低，实现聚类的预训练效果，进而为下游任务提供表情信息的先验知识。在三元对比损失中，除了约束锚点与正负样本的关系外，还需要加入了对正负样本之间关系的约束，即要求锚点与正样本的相似度也要大于正样本与负样本的相似度，同时需要对这一大于关系设置一个最小边界，以确保所有的向量不会集中在同一点。因此，三元对比损失函数最终可以写为：

$$\begin{aligned} L = &\max\left(0, S_{\cos}\left(\boldsymbol{f}_a, \boldsymbol{f}_n\right) - S_{\cos}\left(\boldsymbol{f}_a, \boldsymbol{f}_p\right) + m\arg in\right) \\ &+ \max\left(0, S_{\cos}\left(\boldsymbol{f}_p, \boldsymbol{f}_n\right) - S_{\cos}\left(\boldsymbol{f}_a, \boldsymbol{f}_p\right) + m\arg in\right) \end{aligned} \tag{8}$$

接下来，为了在大规模人脸对比图像数据集上深入挖掘纯粹的表情特征，本研究设计了如图 4 所示的三等级的数据增强方案。

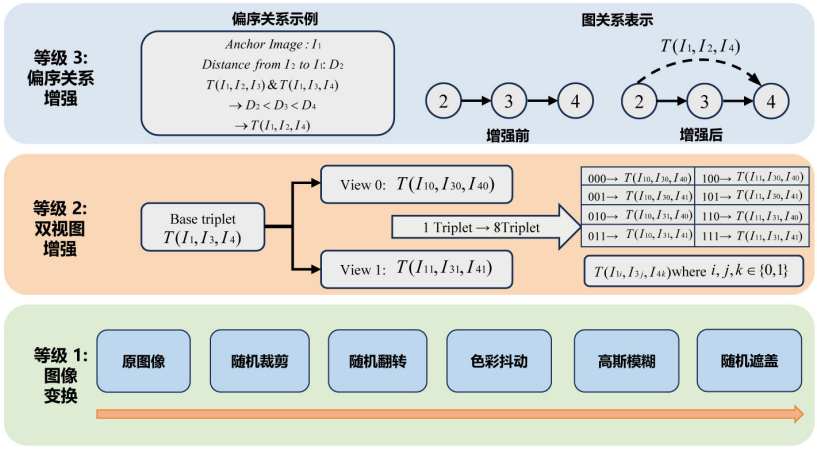

Figure 4. Data augmentation strategies in triplet contrast learning
图 4. 三元组对比学习中的数据增强方案

等级 1 数据增强中，我们使用了随机裁切，随机翻转，色彩增强，高斯模糊，随机遮盖等图像处理过程。

等级 2 数据增强中，我们为每一条包含三张人脸图像的三元对比样本进行两次等级 1 的随机数据增强，得到三张图像的两份不同视图，进而构造出 2^3 = 8 条新的三元对比样本，此方案利用“人脸图像的不同增强视图之间的表情差异为 0”的自然前提，显著扩充了人脸表情对比数据集的等效规模，进一步提高了对比学习效果。

等级 3 数据增强的核心思想是利用公式(7)中表情相似度大小关系的偏序传递性。如图 4 所示，本研究将同一锚点图像的所有三元组样本建模为一个单向图，其中节点为图像，边的方向为正样本图像指向负样本图像，边权固定位 1，代表它们与锚点图像的相似度的大小关系。由公式(7)可知，如果存在三元组(a, b, c)与(a, c, d)，则三元组(a, b, d)一定存在，后者即为等级 3 数据增强中生成的新样本。而在单向图中，这代表从 b 到 c 再到 d 有一条路径的情况下，可以等效于 b 到 d 有一条边权为 1 的边，而如果要挖掘全图中的这类等效边，可以通过在图上运行如算法 1 所示的基于 Bellman-ford 的最短路径算法来发现并量化这种传递关系，进而生成新的三元组样本。根据统计，等级 3 数据增强从原始的约 35 万个三元组标注中，推导出约十万个新的隐含三元组关系。这些新样本为整个数据集提供了更完善的数学约束的同时也丰富了数据集本身的监督信号，进而提高了训练的稳定性以及模型最终性能。

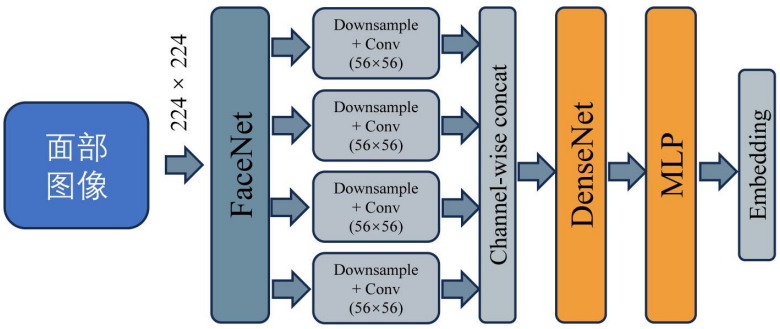

Figure 5. Architecture diagram of MEEN
图 5. MEEN 结构示意图

算法 1 Level-3 数据增强中的 Bellman-Ford 扩展算法

输入：图模型 G = (V, E)，源顶点 s，顶点总数 n
输出：从 s 到所有其他顶点的距离数组 Distance
1. 初始化最短路径权重数组 Distance，Distance[s] = 0，其余 Distance[v] = inf
2. 初始化入队次数数组 count，所有 count[v] = 0
3. 初始化队列 Q，将源顶点 s 压入 Q
4. count[s] = count[s] + 1
5. while Q 非空 do
6. u = Q.pop()
7. for 每一个与 u 相连的邻接点 v，其边权重为 1，do
8 if 1 < Distance[v] then
9. Distance[v] = 1//松弛操作，更新最短路径权值
10. if v 不在队列 Q 中 then
11. if count[v] > n - 1 then
12. 跳过 v，报告检测到环路
13. Q.push(v)
14. count[v] = count[v] + 1
15. 返回数组 Distance，其中所有距离为 1 的点 i 均对应三元组(Anchor, s, i)

3.4. MEEN 表情特征提取网络

本研究使用的多尺寸表情嵌入网络 MEEN 的结构如图 5 所示，MEEN 在经过了 3.3 中的三元对比学习预训练后将作为 CEENet 中的面部特征提取器。MEEN 沿用过去表情嵌入学习中使用的在 Vggface 数据集[27]上预训练的 FaceNet 作为主干网络，如图 5，我们从 FaceNet 的不同卷积层获得提取的 56 * 56，28 * 28，14 * 14，7 * 7 四种尺寸的特征图，并经过下采样以及平滑卷积将它们对齐到 7 * 7 尺寸，并在通道维度进行拼接后送入 DenseNet，网络中所有的参数均可训练。

相比以往的相似工作 FECNet 与 DLN，MEEN 针对表情特征的提取同时考虑了全局特征与局部特征对情绪判别的影响，吸收了过去 CV 领域在目标检测任务上对多尺寸特征融合的成果，在特征图尺寸递减的 FaceNet 网络中增加了类似特征金字塔的结构，将不同尺寸的特征图经过统一尺寸的采样后拼接起来，并且在这个过程中设置平滑卷积确保通道数量的合理性以及特征对齐，进而使得后续的 DenseNet 所接收的特征中同时包含一定的全局与局部特征信息。实验证明经过改进后的 MEEN 在已有的 FEC 数据集上对比 DLN 与 FECNet 等过往工作有着显著的性能提升。

4. 实验

本章将介绍 CEENet 工作当中的具有代表性的实验，以对应前文中的论证过程。首先将展示实验所使用的数据集，第二节将展示实验所使用的软硬件配置，之后将展示在上下文情绪识别任务中与其他已有工作的效果对比，然后展示面部表情对比基准测试中 MEEN 与已有工作的性能对比，最后将对 CEENet 的各部分进行消融实验以验证本工作的有效性。

4.1. 数据集

Emotic 数据集：Emotic [5]数据集是来自 UOC 的情绪识别研究团队在 MSCOCO，framesdb，emodb_small，ade20 四个数据集的基础上，进行人工情感标注而成的数据集，包含了约 2.3 万张图像，每张图片都标注了需要进行情感检测的人的边界框以及其情感种类。

CAER-S 数据集：CAER-S [28]是一个从电视剧中截取的、包含丰富场景信息的多类别情绪识别数据集。它包含了约 7 万张图像，标注了 7 种基本情绪以及人脸边界框。

FEC 数据集：面部表情对比数据集[6]是 Google 研究团队参考了人脸识别任务的训练流程，为了提升人脸表情相关任务的效果，爬取了公开网站上的 155,943 张人脸图像，从中组成了 500,203 个三元组，并安排不同的标注者按照表情相似度进行人工标注，明确每个三元组内的锚点图像，正样本图像与负样本图像。本研究由于 FEC 数据集中的部分公开图片的 URL 已经失效，因此在收集完毕之后，共包括 127,238 张人脸图像以及 356,007 个三元组，规模约为原数据集的 70%，在后续的表格中，我们将使用'Original'来标注在完整数据集上报告的结果。

4.2. 实现细节

数据处理：对 Emotic 与 CAER-S 数据集中的每张图片，使用数据集中附带的 bbox 边界框进行人像裁切，获取人体区域图像，同时使用 opencv 对其进行人脸检测，裁切图像后进行灰度处理，获得人脸区域图像。其中将原始图像与人体区域图像均调整至 224 × 224 像素大小，将人脸图像调整至 48 × 48 像素大小，均使用随机水平翻转与颜色抖动等数据增强方法。对于面部表情对比数据集 FEC，在数据收集阶段，使用多线程爬虫来获取 FEC 数据集中的图像。在三个等级的数据增强之前，每个数据集的图像会统一缩放到 224 * 224 大小，并保持 RGB 通道格式。

软件配置：所有代码实现均使用 Python，其中 FaceNet 与 DenseNet 的实现参考了 Pytorch 库的标准写法。

训练配置：统一在 NVIDIA A800 GPU 上运行本研究的所有实验，通过设置相同的随机种子来确保数据分发与训练过程的一致性。实验中的超参数是控制变量的主要方式，具体设置如下：1) 在 FEC 数据集上的所有的实验均将 batchsize 设置为 128，训练 12 个轮次，使用 SGD 优化器并设置学习率为 0.01，动量设置为 0.9，学习率曲线设置为余弦 1/4 周期衰减，最后设置 5000 步的学习率线性启动以提高训练稳定性；2) 在 Emotic 数据集上的实验统一使用大小为 32 的 batchsize，训练轮次为 10 个 epoch，设置 Adam 优化器的学习率为 5e−4，同时增加指数衰减策略，模型每进行一次前向与反向过程，学习率降至原先的 0.9998 倍。3) 在 CAER-S 数据集上的实验统一使用大小为 1024 的 batchsize，训练轮次为 200 个 epoch，设置 SGD 优化器的学习率为 0.5，动量设置为 0.9，同时使用学习率余弦衰减策略。

4.3. 上下文情绪识别实验

如表 1 所示，我们记录了先前的部分研究在 Emotic 上的结果，另外本研究复现了其中的两种方案，同时把他们与 CEENet 的结果进行对比。其中 Kosti 等人[5]与 CAER-Net [28]均使用了背景 - 身体双编码器网络，Ninh 等人[12]使用了背景 - 身体 - 面部三编码器网络，RRLA [28]则通过身体与背景的特征交互获取额外的情绪信息，EmotionCLIP [24]研究改进 CLIP 预训练过程来学习图像的情感表示，进而应用于上下文情绪识别任务中。CLEF [35]从因果推断角度削弱情绪无关的上下文影响进而提高上下文情绪识别效果。mCFA [36]则研究交叉注意力在上下文融合过程中的作用。对比之下，我们的方法能够在上下文情绪识别领域中处于领先水平。

在 CAER-S 数据集上的实验结果如表 2 所示。CEENet 同样取得了位居前列的成绩。

Table 1. Comparison of metrics for various methods on the Emotic Benchmark
表 1. 各方法在 Emotic 基准测试上的指标对比

方法	平均精度均值(mAP)
CAER-Net [28]	0.2084
Kosti, Ronak, *et al.* [5]	0.2738
Ninh, Quoc-Bao, *et al.* [12]	0.2837
RRLA [29]	0.3241
EmotiCon [13]	0.3203
GCN-CNN [30]	0.2842
DVC-Net [31]	0.2996
EmotionCLIP [24]	0.3291
EMOTIC-Net + CLEF [35]	0.3167
mCFA [36]	0.2877
Kosti, Ronak, *et al.* (Reproduction)	0.2659
Ninh, Quoc-Bao, *et al.* (Reproduction)	0.2743
CEENet (Ours)	**0.3415**

Table 2. Comparison of metrics for various methods on the CAER-S Benchmark
表 2. 各方法在 CAER-S 基准测试上的指标对比

方法	准确率(%)
CAER-Net [28]	73.51
Kosti, Ronak, *et al.* [5]	74.48

续表

GCN-CNN [30]	77.02
SIB-Net [32]	74.56
MHCAN [33]	79.64
GRERN [34]	81.31
RRLA [29]	84.82
EMOTIC-Net + CLEF [35]	77.03
mCFA [36]	85.33
CEENet (Ours)	**88.25**

4.4. 三元表情对比实验

本研究中的主要指标是 FEC 测试集上的准确率，由于我们已有的 FEC 数据集相比原版有一定缺失，因此我们对先前已有的主要方法均进行了复现用以对比 MEEN 的性能，同时也会标出在原版 FEC 数据集上对应模型的准确率。如表 3 所示，MEEN 能够有效提高学习到的表情嵌入的三元准确率，同时在数据样本减少进而造成模型性能下降的情况下，我们的方法仍然可以显著超越过去的方法，仅使用约 70% 的样本的情况下将准确率提高到 86%以上，充分体现了 MEEN 网络以及改进后的数据增强策略对表情嵌入质量的提升效果。

Table 3. Comparison of metrics for various methods on the FEC Benchmark
表 3. 各方法在 FEC 测试集上的指标对比

方法	准确率(%)
AFFNet-CL-P (Original)	53.3
FECNet (Original)	81.8
FECNet	75.3
DLN (Original)	85.4
DLN	82.1
MEEN (Ours)	**86.1**

4.5. 消融实验

CEENet 的主干网络由多个负责不同上下文提取的编码器组成，因此为了验证这些编码器的有效性，我们进行了一系列消融实验。

Table 4. Ablation experiments for different context encoder configurations on the Emotic Benchmark
表 4. 不同上下文编码器配置在 Emotic 基准测试上的消融实验

主干网络(上下文来源)设置				Test Set mAP
背景	身体	面部	视觉嵌入	
✓	✗	✗	✗	0.2621
✗	✓	✗	✗	0.1795
✗	✗	✓	✗	0.1695

续表

✗	✗	✗	✓	0.3022
✓	✓	✗	✗	0.2643
✗	✓	✓	✗	0.1793
✓	✗	✓	✗	0.2664
✓	✗	✗	✓	0.3259
✗	✓	✗	✓	0.3276
✗	✗	✓	✓	0.3354
✓	✓	✓	✗	0.2689
✓	✓	✗	✓	0.3392
✗	✓	✓	✓	0.3402
✓	✗	✓	✓	0.3394
✓	✓	✓	✓	**0.3415**

如表 4 的 15 组实验记录了在屏蔽掉部分 Backbone 的配置下，CEENet 的最终实验结果，可以看出每个编码器对模型的性能都有着一定的提高，同时视觉嵌入编码器对模型情绪识别能力有着比其他编码器更好的贡献，这是由于视觉嵌入编码器能够理解图像背景中更高层级的抽象特征，因此往往能为情绪识别提供关键线索。

Table 5. Ablation experiments for MFA module
表 5. 对 MFA 模块的消融实验

主干网络(上下文来源)设置				Non-MFA	MFA
背景	身体	面部	视觉嵌入		
✓	✓	✗	✗	**0.2676**	0.2643
✓	✓	✓	✗	0.2651	**0.2689**
✓	✓	✓	✓	0.3368	**0.3415**

Table 6. Ablation experiments for data augmentation strategies in triplet contrast learning
表 6. 对三元组对比学习中数据增强方案的消融实验

数据增强选项			准确率(%)
等级 1	等级 2	等级 3	
✗	✗	✗	82.5
✓	✗	✗	83.7
✓	✓	✗	84.6
✓	✓	✓	**86.1**

表 5 记录了屏蔽或使用多流特征注意力模块 MFA 的情况下，CEENet 方法的结果，可以看到编码器使用越多，多流特征注意力对最终结果的提升作用越高。

如表 6 所示，本研究在 FEC 数据集上对 MEEN 所使用的数据增强策略进行消融实验。前 4 行我们从无数据增强开始依次增加方案直到三类数据增强策略齐全，由结果可知，随着数据增强等级的增加，表情嵌入的准确度稳步上升，证明了各项数据增强尤其是等级 3 数据增强的有效性。

MEEN 网络结构上的创新主要来自我们设计的多尺寸特征增强。如表 7 所示，本研究根据多尺寸特征增强的有无以及实现方法(特征图相加还是特征图拼接)，在 FEC 数据集上进行消融实验，结果表明多尺寸特征增强显著提高了表情嵌入的准确度，同时 concat 方案也优于 add 方案，说明拼接方案尽管略微增加了计算量，但能够保留更多全局与局部表情信息，因此能够提升表情嵌入的质量。

Table 7. Ablation experiments for multi-scale feature enhancement of MEEN
表 7. 对于 MEEN 多尺寸特征增强的消融实验

网络设置				准确率(%)
方法	无多尺寸特征增强	多尺寸特征相加	多尺寸特征拼接	
FECNet	✓	✗	✗	75.3
FECNet	✗	✓	✗	76.4
FECNet	✗	✗	✓	76.8
DLN	✓	✗	✗	82.1
DLN	✗	✓	✗	83.0
DLN	✗	✗	✓	83.3
MEEN	✓	✗	✗	83.2
MEEN	✗	✓	✗	85.3
MEEN	✗	✗	✓	**86.1**

5. 结论

本文在情绪识别领域中提出了一种新的上下文情绪识别网络 CEENet，通过引入视觉嵌入扩展了上下文的挖掘深度，同时针对多源上下文特征融合问题提出了自适应注意力模块。在这个过程中，针对 CEENet 中的面部表情特征提取子网络设计了新的表情嵌入模型 MEEN，通过多尺寸特征增强以及经过改进数据增强后的对比学习预训练，显著提高了表情特征质量。CEENet 在上下文情绪识别相关的基准测试中取得了领先的成绩。我们的工作仍在两个方向存在改进空间，一是当前使用的数据集规模限制了模型对比学习能力的上限，未来可以探索在更大规模的数据集上进行训练；二是多流网络的较高的参数量提高了训练成本，未来应当探索更轻量化的网络结构。下一步的工作将围绕上述方向展开。

基金项目

省部级上海市产业协同创新(科技)项目 XTCX-KJ-2023-2-26。

参考文献

[1] Ekman, P. and Friesen, W.V. (1971) Constants across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, **17**, 124-129. https://doi.org/10.1037/h0030377

[2] Matsumoto, D. (1992) More Evidence for the Universality of a Contempt Expression. *Motivation and Emotion*, **16**, 363-368. https://doi.org/10.1007/bf00992972

[3] Bettadapura, V. (2012) Face Expression Recognition and Analysis: The State of the Art. arXiv:1203.6722.

[4] Saxena, A., Khanna, A. and Gupta, D. (2020) Emotion Recognition and Detection Methods: A Comprehensive Survey. *Journal of Artificial Intelligence and Systems*, **2**, 53-79. https://doi.org/10.33969/ais.2020.21005

[5] Kosti, R., Alvarez, J.M., Recasens, A. and Lapedriza, A. (2017) EMOTIC: Emotions in Context Dataset. 2017 *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Honolulu, 21-26 July 2017, 61-69. https://doi.org/10.1109/cvprw.2017.285

[6] Vemulapalli, R. and Agarwala, A. (2019) A Compact Embedding for Facial Expression Similarity. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 5683-5692. https://doi.org/10.1109/cvpr.2019.00583

[7] Zhang, W., Ji, X., Chen, K., Ding, Y. and Fan, C. (2021) Learning a Facial Expression Embedding Disentangled from Identity. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 6759-6768. https://doi.org/10.1109/cvpr46437.2021.00669

[8] Goodfellow, I.J., Erhan, D., Carrier, P.L., *et al*. (2013) Challenges in Representation Learning: A Report on Three Machine Learning Contests. *Neural Information Processing*: *20th International Conference*, Daegu, 3-7 November 2013, 117-124.

[9] Ahmed, F., Bari, A.S.M.H. and Gavrilova, M.L. (2019) Emotion Recognition from Body Movement. *IEEE Access*, **8**, 11761-11781. https://doi.org/10.1109/access.2019.2963113

[10] Cai, Y., Li, X. and Li, J. (2023) Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors*, **23**, Article 2455. https://doi.org/10.3390/s23052455

[11] Mehrabian, A. and Russell, J.A. (1974) An Approach to Environmental Psychology. The MIT Press.

[12] Ninh, Q.B., Nguyen, H.C., Huynh, T., Tran, M. and Le, T. (2023) Multi-Branch Network for Imagery Emotion Prediction. *Proceedings of the* 12*th International Symposium on Information and Communication Technology*, Ho Chi Minh, 7-8 December 2023, 371-378. https://doi.org/10.1145/3628797.3628954

[13] Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A. and Manocha, D. (2020) Emoticon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 14234-14243. https://doi.org/10.1109/cvpr42600.2020.01424

[14] Mittal, T., Bera, A. and Manocha, D. (2021) Multimodal and Context-Aware Emotion Perception Model with Multiplicative Fusion. *IEEE MultiMedia*, **28**, 67-75. https://doi.org/10.1109/mmul.2021.3068387

[15] Radford, A., Kim, J.W., Hallacy, C., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, Virtual, 18-24 July 2021, 8748-8763.

[16] Li, B., Weinberger, K.Q., Belongie, S., *et al*. (2022) Language-Driven Semantic Segmentation. arXiv:2201.03546.

[17] Gu, X., Lin, T.Y., Kuo, W., *et al*. (2021) Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. arXiv:2104.13921.

[18] Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T. and Dubnov, S. (2023) Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. *ICASSP* 2023-2023 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Rhodes Island, 4-10 June 2023, 1-5. https://doi.org/10.1109/icassp49357.2023.10095969

[19] Wang, C., Chai, M., He, M., Chen, D. and Liao, J. (2022) CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 3835-3844. https://doi.org/10.1109/cvpr52688.2022.00381

[20] Li, J., Li, D., Xiong, C., *et al*. (2022) BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. *International Conference on Machine Learning*, Baltimore, 17-23 July 2022, 12888-12900.

[21] Li, J., Li, D., Savarese, S., *et al*. (2023) BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. *International Conference on Machine Learning*, Honolulu, 23-29 July 2023 19730-19742.

[22] Zhang, J., Qu, X., Zhu, T. and Cheng, Y. (2025) CLIP-MoE: Towards Building Mixture of Experts for CLIP with Diversified Multiplet Upcycling. *Proceedings of the* 2025 *Conference on Empirical Methods in Natural Language Processing*, Suzhou, 4-9 November 2025, 5406-5419. https://doi.org/10.18653/v1/2025.emnlp-main.275

[23] Bai, S., Chen, K., Liu, X., *et al*. (2025) Qwen2. 5-vl Technical Report. arXiv:2502.13923.

[24] Zhang, S., Pan, Y. and Wang, J.Z. (2023) Learning Emotion Representations from Verbal and Nonverbal Communication. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 18993-19004. https://doi.org/10.1109/cvpr52729.2023.01821

[25] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[26] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer

Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 10012-10022. https://doi.org/10.1109/iccv48922.2021.00986

[27] Parkhi, O.M., Vedaldi, A. and Zisserman, A. (2015) Deep Face Recognition. In: Xie, X., Jones, M.W. and Tam, G.K.L., Eds., *Procedings of the British Machine Vision Conference* 2015, BMVA Press, 41.1-41.12. https://doi.org/10.5244/c.29.41

[28] Lee, J., Kim, S., Kim, S., Park, J. and Sohn, K. (2019) Context-Aware Emotion Recognition Networks. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October 2019-2 November 2019, 10143-10152. https://doi.org/10.1109/iccv.2019.01024

[29] Li, W., Dong, X. and Wang, Y. (2021) Human Emotion Recognition with Relational Region-Level Analysis. *IEEE Transactions on Affective Computing*, **14**, 650-663. https://doi.org/10.1109/taffc.2021.3064918

[30] Zhang, M., Liang, Y. and Ma, H. (2019) Context-Aware Affective Graph Reasoning for Emotion Recognition. 2019 *IEEE International Conference on Multimedia and Expo* (*ICME*), Shanghai, 8-12 July 2019, 151-156. https://doi.org/10.1109/icme.2019.00034

[31] Qing, L., Wen, H., Chen, H., Jin, R., Cheng, Y. and Peng, Y. (2024) DVC-Net: A New Dual-View Context-Aware Network for Emotion Recognition in the Wild. *Neural Computing and Applications*, **36**, 653-665. https://doi.org/10.1007/s00521-023-09040-8

[32] Li, X., Peng, X. and Ding, C. (2021) Sequential Interactive Biased Network for Context-Aware Emotion Recognition. 2021 *IEEE International Joint Conference on Biometrics* (*IJCB*), Shenzhen, 4-7 August 2021, 1-6. https://doi.org/10.1109/ijcb52358.2021.9484370

[33] Yuan, Y., Lu, F., Cheng, X. and Liu, Y. (2022) Context Based Vision Emotion Recognition in the Wild. 2022 *IEEE* 17th *Conference on Industrial Electronics and Applications* (*ICIEA*), Chengdu, 16-19 December 2022, 479-484. https://doi.org/10.1109/iciea54703.2022.10005917

[34] Gao, Q., Zeng, H., Li, G. and Tong, T. (2021) Graph Reasoning-Based Emotion Recognition Network. *IEEE Access*, **9**, 6488-6497. https://doi.org/10.1109/access.2020.3048693

[35] Yang, D., Yang, K., Li, M., Wang, S., Wang, S. and Zhang, L. (2024) Robust Emotion Recognition in Context Debiasing. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 12447-12457. https://doi.org/10.1109/cvpr52733.2024.01183

[36] Nguyen, H., Tran, N., Nguyen, M., Ta, P. and D. Nguyen, H. (2026) Attention Mechanisms for Context-Aware Emotion Recognition. *CCF Transactions on Pervasive Computing and Interaction*, **8**, 253-269. https://doi.org/10.1007/s42486-025-00216-w