arXiv:2305.03500v1 [cs.CV] 5 May 2023

用于图像情绪识别的高层情境表征

Willamsde Lima Costa Voxar Labs，伯南布哥联邦大学信息学中心，巴西累西腓，
wlc2@cin.ufpe.br

Estefania Talavera Martinez
屯特大学数据管理与生物识别组，荷兰恩斯赫德

Lucas SilvaFigueiredo伯南布哥联邦农村大学贝洛雅尔丁学术单位，巴西贝洛雅尔丁

Veronica Teichrieb
Voxar Labs，伯南布哥联邦大学信息学中心，累西腓，巴西

摘要

情绪识别是对人们感知到的情绪进行分类的任务。以往研究利用各种非语言线索从图像中提取特征，并将其与情绪关联起来。在这些线索中，情境背景对情绪感知尤为关键，因为它会直接影响一个人的情绪。本文提出一种从图像中提取高层情境表征的方法。该模型仅依赖一种线索和一个编码流，将这一表征与情绪关联起来。我们的模型表现可与最先进水平相媲美，在 EMOTIC 数据集上取得了 0.3002 的 mAP，同时能够在消费级硬件上以每秒 ≈ 90 帧的速度运行。总体而言，我们的方法比以往模型更高效，且易于部署，可用于解决与情绪识别相关的现实问题。

1 引言

在人类行为的诸多可能表征中，情绪识别是近几年备受关注的研究课题，这主要得益于能够探索情感计算概念的应用场景迅速发展，例如智能环境和社会机器人学。通过不仅了解用户，也了解其所处的人，系统便能够基于行为知识采取行动，并提出多种干预措施，以改善互动、体验，甚至生活质量。

要识别情绪，系统必须首先解读人类自然发出的情绪信号，有时甚至是无意中发出的 [1]。这些信号通常被称为隐性的，源自一种被定义为非语言交流的沟通形式。研究表明，大量情感信息并非通过言语传达，而是通过这一沟通渠道自然地发送和接收；该渠道涵盖面部表情、肢体语言、语音语调以及其他情绪线索 [2, 3]。这种能力能够带来不受干扰的体验，并感知到未经刻意表现的情绪。

然而，在不受限制的自然场景中，我们的情绪也会受到情境背景信息的影响。例如，一个人在度假时坐在海滩上，享受阳光和海景，更有可能体验到喜悦和快乐等积极情感倾向。相反，一个被困在充斥着噪音污染的交通拥堵中的人，则可能更容易产生沮丧、愤怒或压力等消极情感倾向。因此，在分析情绪时，也应考虑环境刺激。

研究人员多年来一直在提出考虑上下文信息的方法，例如 EMOTIC [4], CAER-Net [5], EmotiCon [6]GLAMOR-Net [7], 和 EmotionRAM [8], ，它们各自提出了新的方法，探讨如何利用上下文，并从不同数据集的图像中提取其表征。我们

假设，尽管过去研究这些低层次的表征特征能够并且已经取得了显著成果，但生成语义层面的高层次描述或许能更好地适应未见数据，从而在测试集上取得更好的结果，并在部署到现实问题中时表现更佳。

在某些场景中，情绪的高层次表征可以为决策者做出明智选择提供有价值的帮助。例如，设想一位城市规划师需要决定应优先改善或翻修城市中的哪些公共公园。为了做出这一决定，规划师需要了解人们身处这些空间时的感受，但不一定需要了解每个人的具体体验。相反，可以将高层次表征以概览的形式呈现，说明与该上下文相关的情绪以及人们对其采取的行动。这样便可以轻松进行比较，而无需处理大量数据。此类方法具有多项优势，例如节省资源。

在本研究中，我们提出了一种方法，用于提取图像的高层情境表征，以实现情绪识别任务。我们通过实验表明，这些具有高度代表性的情境描述能够在 EMOTIC 数据集 [9] 上单独取得与最先进水平相媲美的结果，并且可以轻松地作为情境编码模块集成到完整的情绪识别处理流程中，从而显著提高准确率。我们还证明，我们提出的方法能够实现快速推理，这对于可部署于实际环境中的低功耗边缘设备而言，是一项理想的特性。本研究的贡献如下：

- 我们提出了一种新颖的框架，该框架（i）从图像中提取上下文描述符并构建其高层次表征；（ii）利用图卷积网络（GCN）将这些表征分类为不同的情感类别。
- 我们在广为人知的 EMOTIC 数据集上进行基准测试， [9], 取得了与当前最先进方法相当的准确率。
- 我们讨论了模型对较低计算能力的需求如何使多样化的应用成为可能，从而解决现实世界的问题。

本文其余部分组织如下：第 2 节概述了基于图像的情绪识别相关研究。我们提出的高级上下文描述框架将在第 3 节中介绍。随后，第 4 节和第 5 节分别介绍实验设置并讨论所得结果。最后，我们在第 6 节中得出结论。

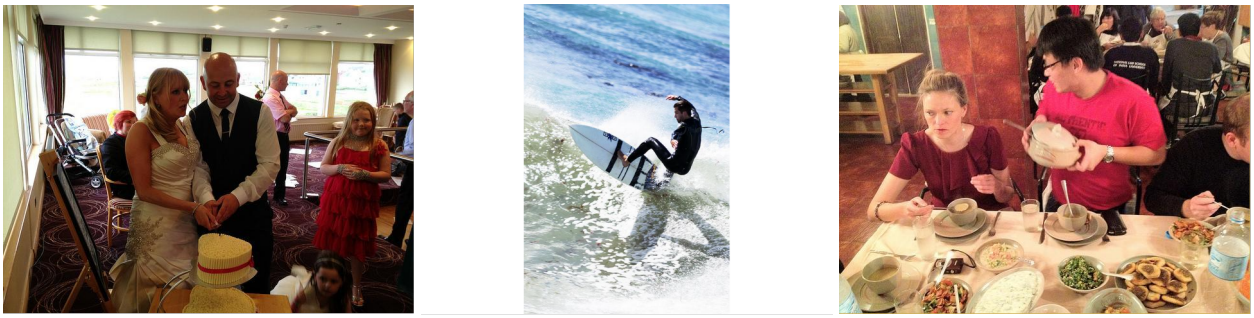
2 文献综述

众所周知，面部表情包含与情绪相关的丰富特征。根据 Barret 的说法，面部之于我们，就如同文字之于页面，而人类能够相对轻松地通过感知解读面部表情 [10]。研究人员自然一直在探索如何提取这些特征，这项任务称为面部表情识别（FER）。然而，在非受控场景中评估 FER 时会遇到一些挑战。面部经常会被部分或完全遮挡，而完全依赖这些信息的技术可能无法取得良好效果。

在这些研究的基础上，研究人员一直在探讨添加其他非语言线索如何改进处理流程。Kosti *et al.* [4] 为这一方法提出了一个基线，该方法会同时考虑人物及其所处的上下文。因此，在发生遮挡的情况下，其他因素也能帮助判断感知到的情绪。他们还发布了 EMOTIC 数据集，用作验证此类方法的基线。EMOTIC 仍被广泛用于相关文献中，以验证基于上下文的技术。

其他技术也被提出，例如 CAER-Net [5] 和 GLAMOR-Net [7], ，它们采用不同的格式来衡量上下文贡献，因此也采用不同方式确定在每种场景中上下文信息的重要程度。然而，上述所有技术在设计上都存在同一局限：未定义哪些内容应被视为上下文。例如， [7] 提出的方法会检测人物的面部，用黑色矩形将其完全遮挡，并使用这张新图像作为上下文的表示。然而，其他身体部位仍然可见，场景中其他人的身体部位也同样可见；随后，这张图像会被输入上下文编码流，该编码流旨在自动从场景中提取特征。然而，这些编码流是否能够在没有先验知识的情况下完成这一任务？

其他方法，例如 EmotiCon [6], 提出，由于上下文的描述性很强，因此有必要使用多个彼此独立且各有所长的流，生成能够与情绪相关联的表征。具体而言，以 EmotiCon 为例，作者提出使用以下情境流：（1）多模态上下文，包含面部特征点和身体关键点；（2）情境背景，通过行人追踪方法遮挡人物后，对背景图像进行处理而提取；以及（3）社会动态上下文，利用深度图计算邻近特征。



Raw caption: A bride and groom cutting a wedding cake
Processed caption: bride groom cut wedding cake

(a)

Raw caption: A man riding a wave on a surfboard in the ocean
Processed caption: ride wave surfboard ocean

(b)

Raw caption: A man and a girl sitting at a table with food
Processed caption: sit table food

(c)

图 1：EMOTIC 中图像的示例及其原始图注和处理后的图注。

在一种较新的方法中，Chen 等人 [11] 提出了结合上下文中不同表征的模型。例如，他们为场景中的每个人分别使用一个深度网络，以计算人与人之间的社会关系，并将其分为亲密关系、非亲密关系和无关系。他们还提出了一个深度推理模块，用于利用多个在局部和全局范围内提取的上下文表征，这些表征涉及场景识别、人体姿态估计等模块。

然而，人类对上下文的感知各不相同 [12, 13]。文献表明，人类会自然地利用自身的内部表征来编码图像中的含义。因此，对我们而言，刻意采取步骤来理解上下文并不自然。相反，大脑会基于我们对这些信息的既有认知，自动将这些刺激归类为正面、中性或负面 [14]。我们的方法不同于前面提到的技术，因为我们对上下文采用了更直接的处理方式，力求基于我们对非语言交流和行为心理学文献的最佳理解，模仿人类的情境表征。

3 高级上下文表示

在本节中，我们描述了从上下文中提取高层表征的方法。鉴于人类描述和理解图像中上下文的方式，我们提出提取图像的高层描述，并将其与语义特征相关联。这种方法模仿人类将语义描述与情绪相关联的方式，以提高可解释性。

3.1 图像描述符提取

高层描述。给定一张输入图像，我们首先希望提取高层图像描述。我们采用基于 Swin-Transformer 架构的 ExpansionNet-v2 [15]，图像描述模型 [16]。我们首先遍历 EMOTIC 数据集，并针对每个样本，将图像输入 ExpansionNet-v2 以生成图注。
随后，我们对原始图注进行处理，以生成优化后的图注。首先，我们从图注中删除停用词。停用词是指语言中常见的词语，例如冠词、介词和代词，但这些词并不具有任何语义意义。因此，考虑到这些词在英语中的高频出现，保留它们只会增加复杂性，而将其删除则有助于获得更具代表性的语料库。我们使用 spaCy[1]，这一用于自然语言处理的公共库。我们还删除了诸如 *man*（男人）、*woman*（女人）、*girl*（女孩）和 *boy*（男孩）等普通名词。此外，我们还应用词形还原，将每个词还原为其词根形式。剩余的词称为 有效词，并将在后续步骤中用于生成数据表征。最后，我们在图 1 中展示了图像、其原始图注以及处理后的图注示例。

共现挖掘。 第二步是生成共现矩阵，用于表示数据集中的标签模式，并在后续通过条件概率加以利用。对数据集中的图像描述进行预处理后，我们会存储这些信息，并统计每种情绪以及每条图像描述中有效词的出现次数，从而得到矩阵 $M_c \in \mathbb{N}^{W \times C}$，其中 W 是语料库中的有效词数量，C 是数据集中的情绪类别数量。因此，$M_{c_{ij}}$ 表示有效词 W_i 出现时，情绪 C_j 出现的次数。我们称这个矩阵为情绪共现矩阵。基于相同的

[1]可在 https://spacy.io/ 获取

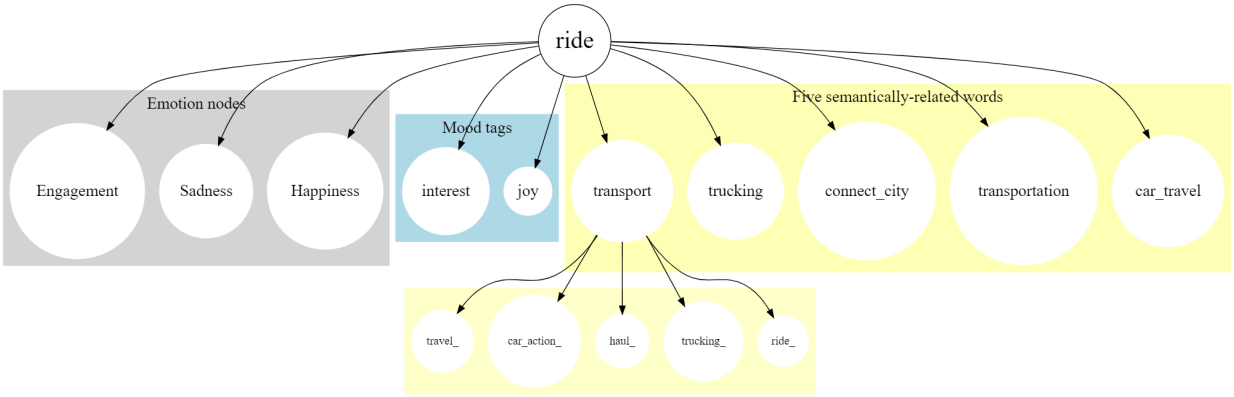


图2：生成图示例。为简洁和便于可视化，在此场景中我们仅考虑一个有效词。对于每个有效词节点，我们为情绪（同样为简洁起见进行了缩减）、情绪标签和语义相关词创建节点。对于每个语义相关词，我们还会查询 SenticNet 并提取其语义相关词，如“transport”节点所示。

基于这一假设，我们还根据有效词的共现情况生成了一个共现矩阵。给定一个大小为 s 的窗口，我们滑动该窗口来捕获有效词的共现情况，从而得到矩阵 $M_w \in \mathbb{N}^{W \times W}$。因此， $M_{W_{ij}}$ 表示有效词 W_i 与有效词 W_j共同出现的次数。

语义描述。 对于每个有效词语 W，我们提取可与情感相关联的语义表征。鉴于 ExpansionNet-v2 是一个在通用上下文中生成图像描述的模型，提取该词语的语义表征将有助于更好地表征该图像描述中的情感含义。

为了提取语义描述，我们采用 SenticNet [17], ，这是一个用于存储与自然语言概念相关的语义、情感语义和极性的知识库。我们查询每个有效单词，并提取以下属性：与该概念相关联的两个情绪标签；愉悦度情感语义，用于表示该词所传达的愉悦或不愉悦感知；极性值，用于表示该词的整体情感倾向；以及在语义上相关的概念。
如果有效词不存在于 SenticNet 知识库中，我们会使用 WordNet [18, 19] 查找同义词。这种方法的优势在于，WordNet 中的词通过同义词集进行分组，而同义词集是由概念或含义相近的同义词组成的集合。因此，通过在 WordNet 中查询某个词，其同义词与该词有很强的关联，且几乎可以确定含义相同。对于每个可能的同义词，我们根据其与有效词的相似度对列表进行排序，然后依次遍历，选取第一个出现在 SenticNet 中的同义词。在极少数情况下，如果有效词及其同义词都不在 SenticNet 中，我们就会从图注中删除该有效词，并在不使用它的情况下继续下一步。

图生成。 凭借这些先验知识（例如，共现和语义表征），我们能够捕捉有效词与情绪之间以及有效词彼此之间的关系。鉴于图特别适合描述结构化数据，我们选择用图来建模这些表征。虽然部分知识是预先学习得到的，但图的定义与构建会按需实时完成。这使该技术能够从未见数据中生成表征。

我们使用深度图学习库[2], ，这是一个与框架无关的图生成与操作库。首先，我们构建一个空图 $G = (V, E)$，其中 V 是节点集合， E 是边集合。在这种情况下，$V = E = \{\emptyset\}$。对于每个有效词语 W，我们首先向图中原本为空的节点集合 V添加一个新节点 V_{W_i}。我们使用 GloVe [20] 获取有效词语的词嵌入，并将该表示用作节点 V_{W_i} 的特征 $X \in \mathbb{R}^{50}$。如果 GloVe 中不存在该有效词语，我们就从均匀分布[-0.01, 0.01]中随机采样该嵌入。我们会保存此表示，以便在该有效词语再次出现时使用。
接下来，我们为数据集中的每个情绪类别 C 添加一个节点 V_C。对于 EMOTIC，由于共有 26 种可能的情绪，因此我们添加 26 个节点，并在有效词语与每种情绪之间建立边 $e = (V_W, V_{C_i})$。我们根据以下公式定义权重 w_e:

[2]可通过 www.dgl.ai 获取

$$w_e = P(C_i|W) = \frac{M_{C_{W,i}}}{\sum M_{C_W}}, \tag{1}$$

其中，有效词节点 V_W 与 i_{th}情绪类别 C_i 之间的边缘端权重 w_e 为 P（$C_i|W$），其计算方式是将有效词 W 与情绪类别 C_i 的共现次数除以从共现矩阵 M 中提取的有效词 W 与所有可能情绪 C 的共现次数之和。

接下来，我们添加与该词的情感语义描述相关的节点。首先，我们添加从 SenticNet 中提取的情绪标签对应的两个节点，并将有效词节点与这两个节点之间的边权重设为该词的愉悦度值。接下来，我们添加通过查询 SenticNet 获得的五个语义相关词对应的五个节点，并以极性值作为权重添加边。

对于与有效词相关的五个语义相关词中的每一个，我们同样查询 SenticNet，并提取其五个语义相关词。第一层词的极性值决定了这些连接之间的边权重。我们推测，通过添加另一层语义关系，可以提取更深层的上下文表征。

我们对图注中的每个有效词 W 执行相同的处理过程。创建所有有效词的节点后，我们在这些节点之间添加边。每条边的权重由词语 $M_{W_{i,j}}$ 的共现次数除以该词出现的总次数得到。最后，为简洁起见，我们在图2中展示一个信息经过简化的图示例。

3.2 用于情绪识别的深层 GCN

鉴于我们构建了用于表示上下文的图，因此使用深层图卷积神经网络进行图分类，进而实现情绪识别。给定一组图 G_1、..... G_N 以及一组情绪类别 $C \in \mathbb{R}^{26}$，我们的目标是根据情绪类别对每个图进行分类。针对这一任务，我们提出采用图同构网络（GIN） [21], , 因为其架构简单，有望在低能耗设备上实现合理的推理速度。

首先，给定一个作为输入的图，我们将该图的特征直接存储在隐藏表示堆栈中，记为 $h0$。之后，我们循环执行一个 GIN 卷积块，该模块包含一个 GIN层、批归一化和 ReLU。在此方法中，我们将该模块迭代五次，生成从 $h1$ 到 $h5$ 的表征。最后，我们遍历隐藏表示，对这些特征进行平均池化并降低其维度。同时，我们保留一个用于类别分类的分支，输出分类标签 C；另一个分支则输出 VAD模型的连续预测结果。VAD模型 [22] 是一种常见的情绪识别方法，它在三维空间中表示情绪，其中每个维度对应情绪体验的不同方面。

VAD模型由效价、唤醒度和支配度构成。效价（V）轴指的是正在体验情绪的个体感受到的愉悦或不愉悦程度。唤醒度（A）表示与情绪相关的能量水平；例如，高唤醒度与更有活力、更兴奋的情绪相关，而低唤醒度则与更平静的情绪相关。最后，支配度（D）反映的是个体对这种情绪状态感受到的控制力或掌控力。研究人员发现，不同情绪往往聚集在VAD空间的不同区域，体现了它们共有的特征及其底层表征。因此，每种情绪都可以表示为三个轴的线性组合 [23]。

因此，我们在训练期间学习类别标签和连续值。我们将损失定义为各项输出的单项损失的加权组合。给定预测 $\hat{y} = (\hat{y}_{cat}, \hat{y}_{cont})$，其中 $\hat{y}_{cat} \in \mathbb{R}^C$ 和 $\hat{y}_{cont} \in \mathbb{R}^3$，我们将该预测的损失定义为 $L = \lambda_{cat} L_{cat} + \lambda_{cont} L_{cont}$，其中 L_{cat} 和 L_{cont} 分别表示各项预测的损失。对于 L_{cat}，实现 EMOTIC [9], 中使用的加权欧几里得损失，其定义如下：

$$L_{2_{cat}}(\hat{y}_{cat}) = \sum_{i=1}^{26} w_i (\hat{y}_{cat_i} - y_{cat_i})^2, \tag{2}$$

其中，$\hat{y}_{cat_i}$ 是 i_{th} 类别的预测值，y_{cat_i} 是其真实标签。权重 w_i 定义为$w_i = {}_{ln}{}^{1}{}_{(c+p_i)}$，其中 p_i 是 i_{th} 类别的概率，c 是用于控制有效值范围的参数。我们还对 L_{cont}采用 L2 损失，其定义如下：

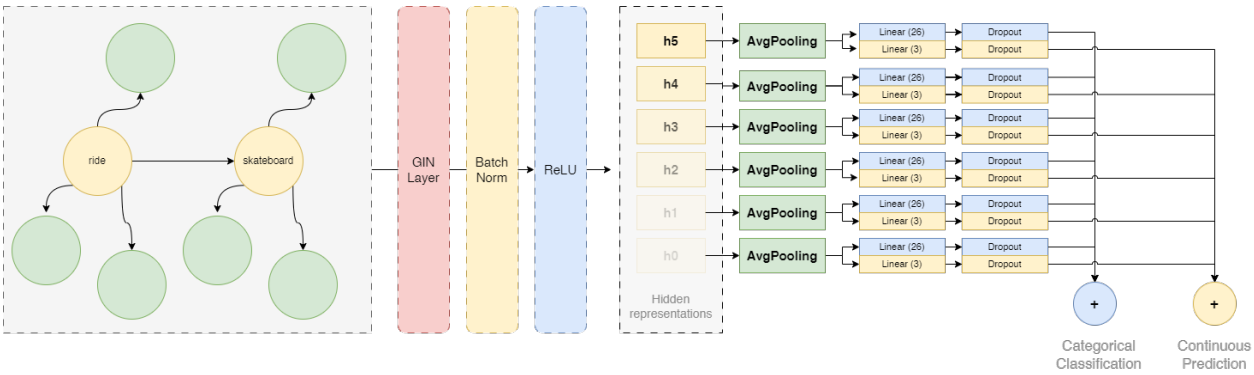


图 3：我们提出的高层情境表征架构。给定一个利用先前学到的知识和图像描述生成的输入图，我们采用图同构网络（GIN） [21]的一种改进方法，将其归入一组情绪类别。鉴于 EMOTIC 还采用连续模型进行标注，我们对处理流程进行了调整，使其生成两个预测结果，并据此计算损失。

$$L_{2_{cont}}(\hat{y}_{cont}) = \sum_{j=1}^{3}(\hat{y}_{cont_j} - y_{cont_j})^2. \tag{3}$$

最后，我们使用上述特征在 EMOTIC 数据集上训练模型。我们使用默认的 PyTorch 数据加载器，并实现了访问图列表的功能，以便在执行过程中将其输入模型。图 3 展示了我们模型的概览。

4 实验

数据集。 我们在情境情绪数据集（EMOTIC） [9], 上开展实验，使用 2019年版本，以便与最先进水平进行直接比较。该数据集侧重于人们处于非摆拍环境中的样本，因此包含展现真实情绪和表情的图像。该数据集还注重图像的上下文可见性和多样性，即图像中包含上下文信息，且样本涵盖多种上下文。作者还从其他成熟的数据集中收集样本，例如 COCO [24] 和 ADE20K [25],，并为这些图像补充了情绪标注。

EMOTIC 使用亚马逊土耳其机器人服务进行标注。训练集中的图像仅标注一次，而测试集和验证集分别由三名和五名标注员标注。据作者称，该数据集包含 23,571 张图像，其中有 34,420 名被标注主体，66% 为男性，34% 为女性；10% 为儿童，7% 为青少年，83% 为成年人。与行为心理学文献的观点不同[26], , EMOTIC 采用了 26 种情感类别，这些类别基于词语关联和词语间的相互依存关系提出，因此形成了词语分组。然而，该数据集缺少中性 类别，这可能会损害标注过程。

与最先进水平的比较。 我们将我们的结果与其他最先进水平的技术进行比较，即 EMOTIC[9], Zhang 的工作 [27], EmotiCon [6], DRM [11], LEKG [11], 这两种方法是 Chen 方法的变体[11], 以及 Yang 的工作 [28]。然而，正如我们稍后在表 1 中所述，这些技术通常建立在多种非语言线索的组合之上。

验证指标。 除了使用当前文献中采用的平均精度均值（mAP）指标进行定量评估 [4, 9, 27, 6, 11], 外，我们还展示了一些示例，以对预测结果进行简要的定性评估。

实现细节。 我们从头开始训练模型，并使用 Adadelta [29]学习参数。在验证集上对多个值进行经验比较后，将批大小设置为 16。我们使用 0.001 的学习率和 0.0004 的权重衰减。

关于实验环境，我们在一台运行 Ubuntu 20.04 LTS 的台式电脑上训练和验证模型。该电脑配备 Intel i7-4790K、32 GB 内存和 12 GB 显存的 NVIDIA RTX 2080 Ti。对于训练

表 1：我们的方法与最先进模型在 EMOTIC 数据集上的定量评估。

技术	mAP	非语言线索的数量
EMOTIC [9]	27.38 20.47*	2（身体和上下文）
Zhang [27]	28.42	1（包含两个流的上下文）
EmotiCon [6]	35.48 26.87*	4（面部描述、身体姿态、上下文和深度映射）
DRM [11]	26.48	5（3条身体描述和2条上下文描述）
LEKG [11]	29.47	2（场景识别和全局上下文）
Yang [28]	**37.73**	5（面部关键点、身体姿态、上下文、关系和智能体-物体交互）
Ours	**30.02**	1（单流上下文）

* Chen 对该工作的复现，报告见 [11]

和测试模型，我们使用 PyTorch 1.12、CUDA 11.3 和 CuDNN 8.3.2。对于推理时间实验，我们还将其与一台运行 Windows 10 专业版的消费级笔记本电脑进行比较，该笔记本电脑配备 16 GB 内存和 6 GB 显存的 NVIDIA GeForce GTX 1060M。

5 结果与讨论

我们在表 1 中将我们的结果与不同方法进行了比较。我们提出的方法优于其他与图相关的方法，例如 Zhang、Liang 和 Ma 的工作 [27]，以及 DRM 和 LEKG；后两者是 Chen等人 [11]方法的两种变体。我们没有将我们的方法与 Chen 的 TEKG 模型进行比较，因为该模型采用局部方法，无法独立扩展到现实世界的问题。尽管 EmotiCon [6] 报告的 mAP 比我们的方法高 5.46 mAP，但根据 Chen 等人 [11], 的说法，他们报告的 35.48 mAP 结果无法复现；复现实验得到的分数为 26.87 mAP，而我们的方法可以高出 3.15 mAP。此外，EmotiCon 使用了四种非语言线索，而我们只采用了一种。Yang 等人 [28] 在该数据集上报告了最高的 mAP，结果为 37.73 mAP，比我们的结果高 7.71 mAP，同时还采用了五种非语言线索。因此，我们证明了即使只使用一种线索，我们的模型仍具备与最先进水平相竞争的性能。Zhang、Liang 和 Ma [27] 同样只使用一种线索，但他们在两个层级上对其进行处理，因此可以被视为使用了两种上下文线索。

所采用的非语言线索数量与模型的推理时间直接相关，而这正是我们希望解决的问题。考虑到现实应用场景，情绪识别模型应当易于部署在边缘端。这样便可设置多个数据采集和处理节点，而无需大幅增加开支或能耗。这是当前在实际环境中部署深度学习模型所面临的两大障碍。不过，在需要使用多种线索的情况下，我们的模型可以充当上下文编码流，即使采用卷积神经网络，也能提取描述信息，并为整体情绪感知做出贡献。

我们对模型进行了消融研究，以评估其在不同配置下的性能。表 2 中呈现的结果表明，对于 GIN，使用求和池化替代当前的池化方法会导致性能下降。此外，我们还将模型的性能与使用由两个 GCN 模块、ReLU 激活函数和分类器组成的简单 GCN 的性能进行了比较。我们的模型优于该简单 GCN，后者取得了更低的 mAP。

我们还通过定性分析对模型进行了评估，如图4所示。对于 EMOTIC 测试集上的每个模型，我们将该图像输入所提出的处理流程，从而生成该图像的类别预测结果。尽管该模型能够

表 2：所提方法的消融研究。

方法	mAP
GIN [21] (平均池化)	0.3002
GIN [21] (求和池化)	0.1715
简单 GCN	0.2505



图4：我们的模型在 EMOTIC数据集上的定性结果。对于每张图像，我们展示了数据集中标注的真实情绪以及网络的预测结果。

此外，我们还使用 VAD模型生成结果。对于人类而言，理解和比较这些值很困难，因为这种表示方式对我们来说并不自然。因此，我们选择仅使用类别值进行定性评估。正如预期，这项实验表明，我们的方法会从上下文中寻找线索来进行情绪预测。在图4a和图4b中，模型能够预测出真实标注中存在的所有类别。在这种情况下，鉴于上下文具有代表性，我们的模型表现良好，能够概括该场景中的情绪。相反，在图4c这类情况下，上下文缺乏代表性，网络无法预测出任何正确的情绪类别。在图4d中，上下文与一组情绪相关，例如积极情绪，但人物感知到的情绪实际上是消极的。在这个例子中，观察人物时，我们能感知到一种与疲倦相关的情绪，这一点也得到了真实标注的证实：*疲劳*。然而，由于模型不会根据上下文观察面部或肢体语言，因此它感知到了错误的情绪。最后，在图4e和图4f中，上下文非常笼统，但模型能够从中提取线索并进行正确分类，至少在一定程度上如此。

最后，我们在不同环境中测试模型，以评估所需的计算能力和推理时间。我们通过将个体预测的批大小设为 1 来执行整个测试流程，并在第 4 节所述的两种环境中进行评估。我们将每个个体预测存入列表，然后计算最小值和平均值。最小值表示推理速度最快的样本所用的时间，而平均值表示模型的平均推理时间。在一台中等配置的深度学习机器上，模型的最短推理时间为 4.0264ms，平均推理时间为 4.1546ms，分别对应 ≈ 248 fps 和 ≈ 240 fps。在一台消费级笔记本电脑上，模型的最短推理时间为 8.9597ms，平均推理时间为 10.3898ms，分别对应 ≈ 111 fps 和 ≈ 96 fps。最后，在同一台消费级笔记本电脑上不使用 CUDA 时，模型的最短推理时间为 13.0021ms，平均推理时间为 17.7365ms，在 Intel Core i7-7700HQ @ 2.80GHz CPU 上分别对应 ≈ 77 fps 和 ≈ 56 fps。
我们没有将推理时间与上文评估的其他方法进行比较，因为它们都没有官方的开源实现。不过，我们可以推断，像 EmotiCon 这样的模型会比我们的方法执行时间更长，因为它使用多种其他深度学习模型来提取和处理特定线索。

6 结论

在这项工作中，我们提出了一种基于图像上下文线索的情绪识别方法。我们提出的方法利用图像描述生成图像的高层次描述。随后，结合数据挖掘和词语语义分析，我们生成并使用作为上下文高层次表征的图。最后，我们采用图分类方法，将这些图归入一组情感类别。尽管其他研究者的方法基于多模态和对上下文的多种描述，但这种简单的方法依然有效，且可与更稳健的模型相媲美。最后，我们展示了该技术速度快、所需计算能力低，因此能够轻松、低成本地部署于多种场景。

未来的工作中，我们计划将这些上下文的高层表征整合到一个多线索模型中，该模型还结合面部和肢体语言，以应对场景中的上下文与人们感知到的情绪不匹配的情况。

致谢

本研究部分经费由巴西高等教育人才培养协调机构（CAPES）资助——资助代码001。

参考文献

[1] Ross Buck. 非语言交流：自发性与象征性方面。《美国行为科学家》，31(3):341–354, 1988.[2] Philipp V Rouast, Marc TP Adam 和 Raymond Chiong. 人类情感识别的深度学习：见解与新进展。《IEEE情感计算汇刊》，12(2):524–543, 2019.[3] Dipika S Patel. 肢体语言：一种有效的交流工具。《IUP英语研究期刊》，第2页，2014.[4] Ronak Kosti, Jose M Alvarez, Adria Recasens 和 Agata Lapedriza. 语境中的情绪识别。《IEEE计算机视觉与模式识别会议 (CVPR)》，第1667–1675页，2017.[5] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park 和 Kwanghoon Sohn. 语境感知情绪识别网络。《IEEE国际计算机视觉会议 (ICCV)》，第10143–10152页，2019.[6] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera 和 Dinesh Manocha. Emoticon：基于弗雷格原则的语境感知多模态情绪识别。2020年6月.[7] Nhat Le, Khanh Nguyen, Anh Nguyen 和 Bac Le. 用于情绪识别的全局-局部注意力。《神经计算与应用》，第1–15页，2021.[8] Willams Costa, David Macêdo, Cleber Zanchettin, Estefanía Talavera, Lucas Silva Figueiredo 和 Veronica Teichrieb. 一种用于人类情绪识别的快速多线索融合方法。《SSRN预印本 4255748》，2022.[9] Ronak Kosti, Jose M Alvarez, Adria Recasens 和 Agata Lapedriza. 使用 EMOTIC 数据集的基于语境的情绪识别。《IEEE模式分析与机器智能汇刊》，42(11):2755–2766, 2019.[10] Lisa Feldman Barrett, Kristen A Lindquist 和 Maria Gendron. 作为情绪感知语境的语言。《认知科学趋势》，11(8):327–332, 2007.[11] Jing Chen, Tao Yang, Ziqiang Huang, Kejun Wang, Meichen Liu 和 Chunyan Lyu. 将结构化情绪常识知识和人际关系融入语境感知情绪识别。《应用智能》，53(4):4201–4217, 2023.[12] Lisa Feldman Barrett 和 Elizabeth A Kensinger. 情绪感知过程中通常会编码上下文。《心理科学》，21(4):595–599, 2010.[13] Lisa Feldman Barrett, Batja Mesquita 和 Maria Gendron. 情绪感知中的语境。《心理科学当前方向》，20(5):286–290, 2011.[14] M Carmen Pastor, Margaret M Bradley, Andreas Löw, Francesco Versace, Javier Moltó 和 Peter J Lang. 情感图片感知：情绪、语境与晚期正电位。《脑研究》，1189:145–151, 2008.[15] Jia Cheng Hu, Roberto Cavicchioli 和 Alessandro Capotondi. Expansionnet v2：用于图像描述快速端到端训练的块静态扩展。《arXiv预印本 arXiv:2208.06551》，2022.[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin 和 Baining Guo. Swin Transformer：使用移位窗口的分层视觉变换器。载于《IEEE/CVF国际计算机视觉会议 (CVPR)》，第10012–10022页，2021.[17] Erik Cambria 和 Amir Hussain.《Sentic Computing：一种基于常识的概念级情感分析框架》。Springer Cham，2015.[18] George A Miller. WordNet：英语词汇数据库。《ACM通讯》，38(11):39–41, 1995.[19] George A Miller.《WordNet：电子词汇数据库》。麻省理工学院出版社，1998.[20] Jeffrey Pennington, Richard Socher 和 Christopher D Manning. GloVe：词表示全局向量。载于《自然语言处理实证方法会议 (EMNLP)》，第1532–1543页，2014.

[21] Keyulu Xu、Weihua Hu、Jure Leskovec 和 Stefanie Jegelka。图神经网络的能力有多强？*arXiv*预印本 *arXiv:1810.00826*，2018年。[22] Albert Mehrabian。一般心理学理论的基本维度：对人格、社会、环境和发展研究的启示。1980年。[23] Agata Kołakowska、Wioleta Szwoch 和 Mariusz Szwoch。基于智能手机传感器采集数据的情绪识别方法综述。*Sensors*，20(21):6367，2020年。[24] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。Microsoft coco：上下文中的常见对象。载于欧洲计算机视觉会议，第740–755页。Springer，2014年。[25] Bolei Zhou、Hang Zhao、Xavier Puig、Tete Xiao、Sanja Fidler、Adela Barriuso 和 Antonio Torralba。通过ADE20K数据集实现对场景的语义理解。*InternationalJournal of ComputerVision*，127:302–321，2019年。[26] Paul Ekman。关于基本情绪的论述。*Cognition and Emotion*，6(3-4):169–200，1992年。[27] Minghui Zhang、Yumeng Liang 和 Huadong Ma。用于情绪识别的上下文感知情感图推理。载于*IEEE*多媒体与博览国际会议 *(ICME)*，第151–156页。IEEE，2019年。[28] Dingkang Yang、Shuai Huang、Shunli Wang、Yang Liu、Peng Zhai、Liuzhen Su、Mingcheng Li 和 Lihua Zhang。面向多重上下文感知的情绪识别。载于计算机视觉–*ECCV 2022*：第*17*届欧洲计算机视觉会议，特拉维夫，以色列，*2022*年*10*月*23*日至*27*日，论文集，第*XXXVII*部分，第144–162页。Springer，2022年。[29] Matthew D Zeiler。Adadelta：一种自适应学习率方法。*arXiv*预印本 *arXiv:1212.5701*，2012年。