# 基于注意力机制多尺度网络的自然场景情绪识别

晋儒龙，卿粼波，文虹茜
(四川大学电子信息学院，成都 610065)

**摘　要**：情绪识别作为计算机视觉的一项基本课题已经取得很大进展，然而在无约束自然场景中的情绪识别仍具挑战性.现有方法主要是利用人脸、姿态以及场景信息识别情绪，但是忽略了人物个体在场景中的不确定性，以及不能很好地挖掘场景中的情绪线索.针对现有研究存在的问题，提出了基于人物与场景线索的双分支网络结构，两个分支独立学习，通过早期融合得到情绪分类结果.对于人物在场景中的不确定性，引入身体注意力机制预判人物情绪置信度进而获得人体的特征表示，场景中引入空间注意力机制和特征金字塔以便充分获得场景中不同粒度的情绪线索.实验结果表明，此方法有效融合人物与场景信息，在 EMOTIC 数据集下能够明显提高情绪识别率.

**关键词**：情绪识别；场景理解；注意力机制；特征金字塔
**中图分类号**：TP391.4　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2022.012003

# Emotion recognition of the natural scenes based on attention mechanism and multi-scale network

*JIN Ru-Long*, *QING Lin-Bo*, *WEN Hong-Qian*
(College of Electronic and Information Engineering, Sichuan University, Chengdu 610065, China)

**Abstract**: Emotion recognition as a basic subject in computer vision has made tremendous progress, yet emotion recognition in natural, unconstrained environments is still challenging. Existing methods mainly use face, posture, and contextual information to recognize emotions, but these methods ignore the uncertainty of individuals in the context, and do not tap the emotional cues in the scene well. Aiming at the problems in existing research, a dual-branch network structure based on individual's body and contextual information is proposed, in which two branches are learning independently to get the result of emotion classification through early fusion. For uncertainties of people in the scenes, the body gesture attention mechanism is utilized to estimate the confidence coefficient the body's feature representation is extracted. For context branch, spatial attention mechanism and feature pyramid network are employed to fully obtain the emotional cues of different granularities in the scene. The experiment results demonstrated that the effectiveness of the proposed method in the EMOTIC dataset.

**Keywords**: Emotion recognition; Scene understanding; Attention mechanism; Feature pyramid

**收稿日期**：2020-12-04
**基金项目**：国家自然科学基金(61871278)；四川省科技计划项目(2018HH0143)
**作者简介**：晋儒龙(1997－)，男，贵州六盘水人，硕士研究生，研究方向为计算机视觉. E-mail：jinrulong@stu.scu.edu.cn
**通讯作者**：卿粼波. Email：qing_lb@scu.edu.cn

# 1 引　言

情绪识别是计算机视觉的一项基本任务,它是情感计算的一部分,旨在识别出某个体的感受与状态,例如高兴、悲伤、厌恶和惊喜等.情绪识别技术用途广泛,目前已经在人机交互[1]、安防[2]和医疗健康[3]等领域有所应用,然而自然场景中的情绪识别存在识别难度大等问题仍具挑战性.得益于深度学习近几年的快速发展,基于卷积神经网络(Convolutional Neural Network, CNN)的方法已经成为各种先进模型的基础.

情绪表达的途径多种多样,语音、文本、生理电信号以及图像[4-7]是情绪识别的常见方式.在自然环境中,语音和文本数据难以采集,生理电信号的采集会对研究对象的情绪产生干预,因此基于视觉信息仍是主要的情绪识别方法.关于面部情绪识别,无论是传统的手工提取特征,还是深度学习方法,多数都是关注面部特征,因其能够提供最明显直观的情感状态.普遍使用的方法是面部动作编码系统(Facial Action Coding System, FACS)[8],核心思想是将面部定义为多个运动单元(Action Unit),然后根据不同运动单元的组合编码为 6 种基本表情(快乐、悲伤、恐惧、惊讶、愤怒和嫉妒).由于深度 CNN 网络的快速发展,运动单元从手工设计转变为自动识别,例如 Jain 等[9]提出使用 CNN 进行特征提取和情绪分类.但是人脸在自然场景中存在光照不均匀、遮挡和拍摄角度等问题,导致难以准确识别其情绪状态.

关于姿态情绪识别,Nicolaou 等[10]提出一种面部结合肩部运动信息的情绪识别方法,Schindler 等[11]使用身体姿态在约束条件下识别 6 种基本情绪.Dael 等[12]发现身体的动作和姿态不仅能反映情绪强度,还能得到具体的情绪类别.然而,同一种姿态或行为在不同语境中表达的可能是不同的情绪状态.例如,在家中看电脑和在办公室看电脑是同一种行为,综合考虑其姿势,衣着以及环境会得到情绪状态不同的结论.

最后是基于场景的情绪识别,Mou 等[13]通过融合人脸,身体以及场景信息进行群体的情绪识别,但基于场景信息的个体情绪识别很少被研究.为了更好地研究基于场景的情绪识别 Kosti 等[14]提出了 EMOTIC (EMOTions In Context database)数据集,并且基于该数据集设计了一个双通道的基准网络结构,分别用于提取人物特征和场景特征.在此基础上,Zhang 等[15]利用 Region Proposal Network (RPN)网络提取场景元素作为节点构建情感图进行情绪识别.Bendjoudi 等[16]在双通道的基准网络上提出多任务损失函数改进模型的训练过程.虽然上述方法都利用了场景信息,但是自然场景中的情绪线索有大小远近之分,简单地对场景信息提取特征,并不能有效利用场景中的情感线索.

为了改善上述问题,本文提出了一种基于注意力机制的多尺度情绪识别网络模型.此网络由人物分支与场景分支组成.针对人物个体在自然场景中存在的不确定性问题,人物分支设计一种身体注意力机制用来预判个体情绪的置信度,并且作用于人物的特征,从而抑制相应的不确定性.针对场景情绪线索探索不充分的问题,场景分支设计了全局-局部的网络结构.对于全局信息,利用空间注意力机制获取场景中的全局信息.对于局部信息,利用空间金字塔能够捕获不同粒度信息的能力,将场景中多种尺度的情感线索进行融合增强,从而获得更加丰富的场景特征表示.最后早期融合双分支的特征向量,得到最终的情绪分类结果.本文的主要贡献如下:(1) 提出一种基于注意力机制与多尺度的网络,充分捕获人物与场景各自的情感线索,最后融合二者之间的关系,推理出人物在自然场景中的情绪类别;(2) 在 EMOTIC 数据集进行广泛的实验,实验结果证明了提出模型的有效性.

# 2 模型结构

现有方法在探索人物与场景线索时,只是简单地提取特征,然后进行融合并进行情绪分类,并未关注人物在场景中的不确定性,以及场景信息的复杂性.针对以上问题,设计一种基于注意力机制的多尺度网络情绪识别模型,系统框架如图 1 所示.对于人物个体,提取特征的同时使用注意力机制学习当前人物情绪的置信度;对于场景,使用特征金字塔提取不同尺度的特征图,其中高阶语义信息使用空间注意力机制学习场景中的主要信息,最后融合双分支网络获得情绪分类的结果.

## 2.1 人物分支

在图像中,人物个体能够直观地描述情绪状态,因此建立基于人体的 CNN 网络结构.为避免过拟合以及增强模型泛化性能,使用 Image-Net 数据集下预训练的 ResNet-50 模型进行微调.根据 Bounding Box 裁剪出人物区域,作为网络的输入

$I_B \in \mathbb{R}^{3\times W\times H}$，通过 ResNet-50 得到的特征向量记作 $X_B \in \mathbb{R}^{1\times d}$，其中 $d$ 表示情绪类别数. 其前向传播如式(1)所示

$$X_B = F(I_B; W_B) \tag{1}$$

其中，$W_B$ 表示网络权重. 考虑到图像中人物的遮挡以及人物在图像中是否占主导地位的因素，加入注意力机制预判当前人物对情绪识别的置信度. 该注意力机制有两点值得注意：(1) 位置不同于传统的注意力机制. 不是位于特征图之后，而是直接置于特征提取之前，这样可以有效地预判当前人物的情绪置信度；(2) 结构不同于 Squeeze-and-Excitation[17] 模型. 首先使用全局平均池化得到 1 ×1 的卷积核，再通过两个卷积层得到权重 λ，最后与 $X_B$ 点乘，得到基于人物 CNN 的分类结果如式(2)所示.

$$f_B = \lambda \otimes X_B \tag{2}$$

其中，⊗表示按位置相乘. 部分判决结果如图 2a 所示.

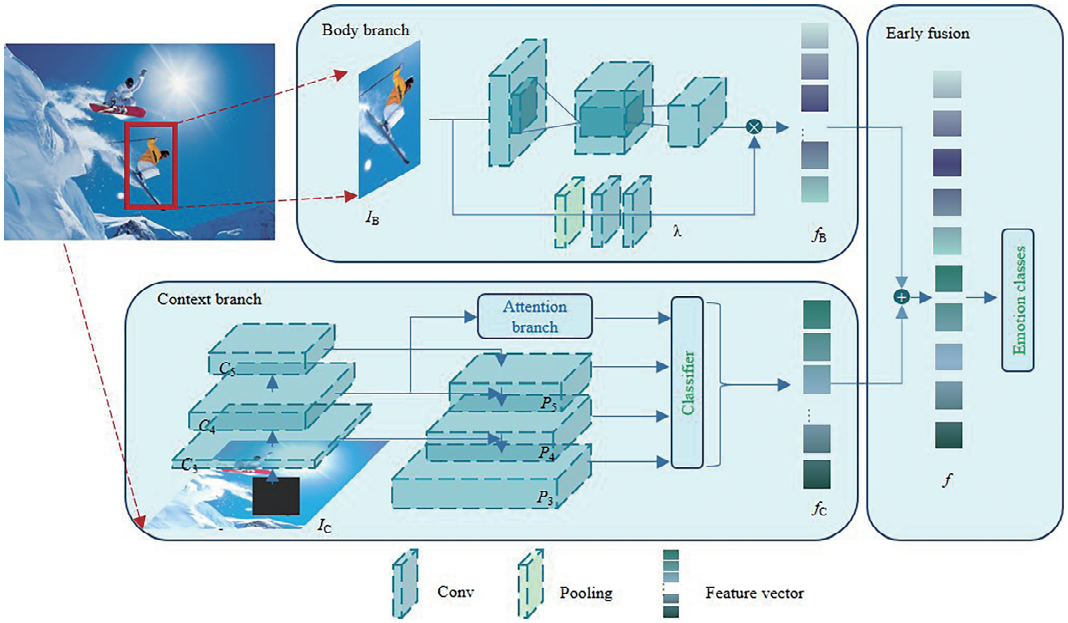


图 1 基于注意力机制的多尺度网络情绪识别框架

Fig. 1　The framework of attention mechanism and multi-scale network based emotion recognition

## 2.2　场景分支

文献[13-15，18]的研究已表明场景信息能够很好地辅助情绪识别，因此搭建基于场景的 CNN 网络结构. 为了防止与人物特征重复提取，对场景中主要人物增加掩模，如式(3)所示，对于场景图像 $I_C \in \mathbb{R}^{3\times W\times H}$ 有

$$I_C = \begin{cases} I(i,j), & \text{if } \notin bbox_{I_B} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

其中，$bbox_{I_B}$ 表示主要人物所在区域. 使用特征金字塔(Feature Pyramid Networks，FPN)[19] 处理场景细节信息. FPN 常用于多尺度目标检测，它能够在增加少量计算量的前提下融合低分辨率语义信息较强的特征图和高分辨率语义信息较弱但空间信息丰富的特征图，在下采样过程中有效地增强局部细节特征. FPN 分为自底向上和自上而下两个过程，在自底向上的过程中，采用预训练的 ResNet-18 模型作为特征提取网络，ResNet 拥有 4 个残差块，为避免内存占用以及过拟合问题，使用最后 3 个残差块的输出构建 FPN，记作 $C=\{C_3, C_4, C_5\}$，分别对应 $I_C$ 的{8，16，32}下采样倍数；在自顶向下的过程中，采用两倍最近邻插值对 $\{C_3, C_4, C_5\}$ 上采样，然后与其下一层的特征图进行对应位置的相加，得到对应的特征金字塔 $P=\{P_3, P_4, P_5\}$，$C$ 与 $P$ 拥有相同的尺寸. 由于 $P$ 共享同一个分类器，所以在分类前通过 1 ×1 卷积修正所有特征图的通道为 256 维，分类器由两个卷积层和全局平均池化构成，输出分类结果为

$$f_{C1} = f_3 \bowtie f_4 \bowtie f_5 \tag{4}$$

由于只关注场景中对情绪识别有帮助的部分，因此引入空间注意力机制，对此使用 Attention Branch Network[20]，与 FPN 自底向上过程共享网络权重，该网络能够有效地识别定位图像中主要的区域，其输出记作 $f_{C2}$. 场景分支的分类结果由 $f_{C1}$ 和 $f_{C2}$ 构成.

### 2.3 模型融合

为融合人物分支和场景分支的特征向量,使用早期融合在通道维数连接

$$f=\text{concatnate}[f_B,f_{C1},f_{C2}] \tag{5}$$

然后通过一个全连接层对特征向量 $f\in\mathbb{R}^{1\times4\times d}$ 进行分类,再通过 Softmax 归一化到[0, 1]区间.

## 3 实验与分析

### 3.1 实验数据

本文基于 EMOTIC 数据集[14]进行实验,该数据集图片来源于 MSCOCO、Ade20K 和网络下载等 3 部分. 共包含 23 571 张图片,标注了 34 320 个人物. 标注信息包含 26 类情绪,每个人物至少拥有一种情绪标签. 其中 70%用于训练,10%用于验证,20%用于测试.

### 3.2 实验设置

本文在 Ubuntu16.04 系统使用 Pytorch 框架进行实验,GPU 为 NVIDIA GeForce GTX2080,内存为 11 GB,模型参数的优化使用 Adam 优化器,初始学习率为 1e-4 并按照余弦方式下降,训练轮数为 70 次,批次为 32,使用 MultiLabelSoftMarginLoss 函数进行误差反向传播. $I_B$ 和 $I_C$ 缩放为 224 ×224,使用水平翻转,改变对比度、亮度和饱和度进行数据增强.

### 3.3 实验分析

沿用文献[14]使用的 mAP(mean Average Precision)作为评价指标以便客观评价模型性能. 实验对比了 EMOTIC 数据集的基准方法[18],Bendjoudi 等[16]提出的方法以及 Zhang 等[15]提出的方法. 实验结果如表 1 所示,从表 1 可以发现自然场景中的复杂情绪识别任务挑战较大. 文献[16]在基准模型[18]的基础上对损失函数进行改进,获得了一定的性能提升. 先进模型[15]利用目标检测算法进一步提取场景线索,其性能的提升也说明有效利用场景线索可以辅助情绪识别.

本文模型利用多尺度信息以及空间注意力机制探索不同粒度的场景信息,相比单阶段方法[16,18]分别提升了 2.27%和 1.32%. 相比于先进模型[15]使用双阶段的训练策略(先单独检测自然场景的线索,然后依赖图神经网络构建情感计算图),我们的模型可以实现端到端的训练以及计算量的减少,并且 mAP 提升了 1.23%,表明了本文提出模型的优越性.

表 1　EMOTIC 测试集下的 AP 和 mAP

Tab. 1　Quantitative evaluation of EMOTIC in comparison on average precision and mean average precision

| 测试集 | AP/% | | | |
|---|---|---|---|---|
| Category | 文献[18] | 文献[16] | 文献[15] | 本文 |
| Affection | 27.85 | 31.92 | 46.89 | 35.09 |
| Anger | 9.49 | 13.94 | 10.87 | 13.93 |
| Annoyance | 14.06 | 17.42 | 11.23 | 18.81 |
| Anticipation | 58.64 | 57.73 | 62.64 | 57.08 |
| Aversion | 7.48 | 8.18 | 5.93 | 10.3 |
| Confidence | 78.35 | 75.29 | 72.49 | 76.28 |
| Disapproval | 14.97 | 14.88 | 11.28 | 19.64 |
| Disconnection | 21.32 | 28.32 | 26.91 | 29.94 |
| Disquietment | 16.89 | 19.72 | 16.94 | 20.27 |
| Doubt/Confusion | 29.63 | 23.11 | 18.68 | 21.33 |
| Embarrassment | 3.18 | 2.84 | 1.94 | 2.83 |
| Engagement | 87.53 | 85.83 | 88.56 | 87.08 |
| Esteem | 17.73 | 16.72 | 13.33 | 16.24 |
| Excitement | 77.16 | 70.43 | 71.89 | 71.49 |
| Fatigue | 9.7 | 14.43 | 13.26 | 13.75 |
| Fear | 14.14 | 8.27 | 4.21 | 8.9 |
| Happiness | 58.26 | 76.61 | 73.26 | 78.82 |
| Pain | 8.94 | 9.38 | 6.52 | 11.59 |
| Peace | 21.56 | 24.31 | 32.85 | 25.21 |
| Pleasure | 45.46 | 46.89 | 57.46 | 47.67 |
| Sadness | 19.66 | 23.94 | 25.42 | 28.87 |
| Sensitivity | 9.28 | 6.28 | 5.99 | 11.47 |
| Suffering | 18.84 | 26.24 | 23.39 | 29.92 |
| Surprise | 18.81 | 10.07 | 9.02 | 10.57 |
| Sympathy | 14.71 | 13.98 | 17.53 | 14.25 |
| Yearning | 8.34 | 9.71 | 10.55 | 9.55 |
| mAP/% | 27.38 | 28.33 | 28.42 | 29.65 |

值得注意的是,在数据较少的类别 Annoyance(2%), Aversion(1%), Disapproval(2%), Pain(1%),Sadness(2%),AP 值均优于其他方法,其中 Disapproval 提升最多为 4.67%. 说明当数据较少时,本文模型仍能有效学习到自然场景中的情绪线索. 在数据较多的类别 Confidence(23%),Engagement(50%),Happiness(26%),相比其他方法,提升幅度有限. 主要是因为场景分支未使用较深层数的骨干网络,这使得我们的模型能够在数据少的类别实现明显的性能提升. 但随着网络的加深,容易导致模型过拟合,进而降低泛化性能. 详细类别分布见文献[14]. 整体来说,所提出模型在多数类别的 AP 均获得了提升,mAP 也达到了最优结果.

为了更准确地衡量本文模型的有效性,设计消

融实验对比分析身体注意力机制(w/BA)，空间注意力机制(w/CA)以及特征金字塔(w/FPN)等 3 个组件的性能，实验结果如表 2 所示. 可见三者组合使用可以获得最优的性能，三者单独使用也优于其他方法. 其中 BA 用于预判人物在场景中的情绪置信度，同时抑制人物个体的不确定性，性能相比先进方法提升了 0.67%. CA 用来捕获全局场景信息，提升最多为 1.85%，用来提取局部场景信息的 FPN 也获得了可比的性能提升. 消融实验结果表明该模型使用的 3 个模块能够充分利用人物信息和场景中的全局-局部信息，从而有效提高情绪识别效果.

除了实验数据分析，我们也对部分测试集进行可视化分析，如图 2 所示. 一方面是人物分支的情绪权重 λ，如图 2(a)所示，当人物在图片中清晰可见时，其权重较大；当人物受到分辨率，拍摄角度等影响，通过人物本身难以识别其情绪状态，对应的情绪权重也相应减少并弱化人物对情绪识别的影响. 另一方面是场景分支的空间注意力分布，如图 2 (c)和(d)所示，对人物增加掩模后，场景分支将注意力从人物本身转移到关注场景本身，这样可以有效地避免人物分支与场景分支学习到重复的特征. 表 3 的实验结果也表明对人物增加掩模后(w/masking)性能有所提升.

**表 2　基于注意力机制的多尺度网络消融实验**

Tab. 2　Ablation studies for proposed method

| w/BA | w/CA | w/FPN | mAP/% |
|---|---|---|---|
| ✓ | | | 29.09 |
| | ✓ | | 29.23 |
| | | ✓ | 28.98 |
| ✓ | ✓ | | 29.35 |
| ✓ | ✓ | ✓ | 29.65 |

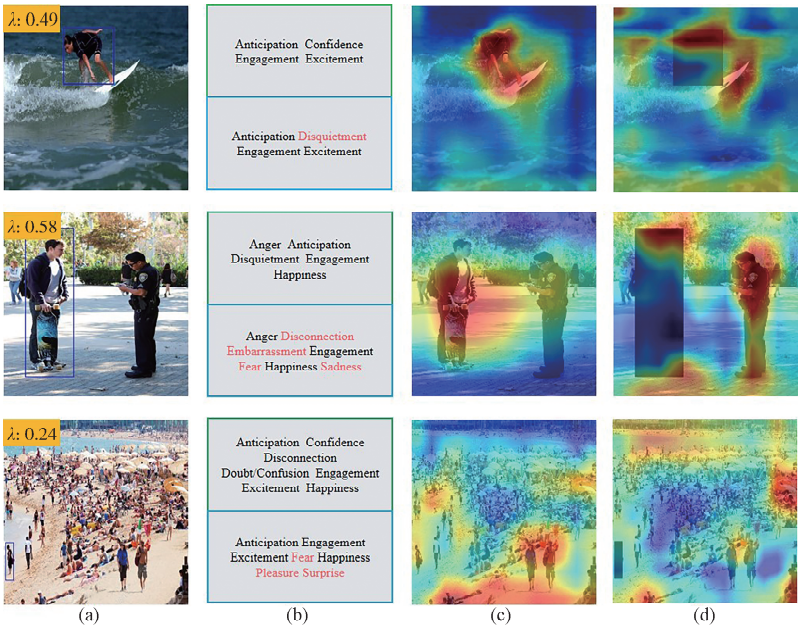


图 2　情绪识别结果可视化

(a) 原始图像；(b) 人物情绪标签，其中绿框为真值，蓝框为预测值；(c)(d) 分别为原始图像和 $I_C$ 训练得到的空间注意力分布

Fig. 2　Visualization of emotion recognition results

(a) Original image; (b) multi-label, which ground truth in green box and prediction in blue box; (c) and (d) results of without hiding the body and with hiding the body during training respectively

**表 3　人物增加掩模性能对比**

Tab. 3　Quantitative evaluation of with/without masking.

| Methods | mAP/% |
|---|---|
| w/o masking | 29.22 |
| w/masking | 29.65 |

## 4　结　论

本文研究了基于人物与场景线索的自然场景情绪识别问题，提出了基于注意力机制的多尺度情绪识别网络结构，在完全缺乏人脸信息的真实场景中，实现了对 26 类复杂情绪的基本识别. 网络结构由人物分支与场景分支组成，针对人物分支设计的身体注意力机制能够有效预判当前人物对情绪识别的置信度，针对场景分支，融合空间注意力机制和特征金字塔可以进一步探索场景中的全局-局部情绪线索. 在 EMOTIC 数据集上进行多个实验以

评估该方法的识别性能.与相关方法比较,实验结果验证了该模型的有效性.虽然本文方法在识别精度上有较好的结果,但仍然有进一步的提升空间,主要原因是在对人物分支以及数据集不平衡的研究有限,在后续研究中,会考虑融合行为识别和改进训练策略等方式,提升算法识别的精度.

**参考文献:**


[1] Hortensius R, Hekele F, Cross E S. The perception of emotion in artificial agents [J]. IEEE T Cogn Develop, 2018, 10: 852.

[2] Clavel C, Vasilescu I, Devillers L, *et al*. Fear-type emotion recognition for future audio-based surveillance systems [J]. Speech Commun, 2008, 50: 487.

[3] Anderson C L, Agarwal R. The digitization of healthcare: boundary risks, emotion, and consumer willingness to disclose personal health information [J]. Inf Syst Res, 2011, 22: 469.

[4] Zhao J, Mao X, Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks [J]. Biomed Signal Proces, 2019, 47: 312.

[5] 陈黎, 刘雨欣, 周耘立, 等. 融合表情符号图像特征学习的微博情感分类[J]. 四川大学学报:自然科学版, 2021, 58: 74.

[6] 赵容梅, 熊熙, 琚生根, 等. 基于混合神经网络的中文隐式情感分析 [J]. 四川大学学报:自然科学版, 2020, 57: 264.

[7] 陈文绪, 薛晓军, 许江淳, 等.多级细节信息融合的人脸表情识别[J]. 重庆邮电大学学报:自然科学版, 2021, 33: 304.

[8] Friesen E, Ekman P. Facial action coding system: a technique for the measurement of facial movement [J]. Palo Alto, 1978, 3: 5.

[9] Jain D K, Pourya S, Paramjit S. Extended deep neural network for facial emotion recognition [J]. Pattern Recogn Lett, 2019, 120: 69.

[10] Nicolaou M A, Gunes H, Pantic M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space [J]. IEEE T Affect Comput, 2011, 2: 92.

[11] Schindler K, Van Gool L, De Gelder B. Recognizing emotions expressed by body pose: a biologically inspired neural model [J]. Neural Networks, 2008, 21: 1238.

[12] Dael N, Mortillaro M, Scherer K R. Emotion expression in body action and posture [J]. Emotion, 2012, 12: 1085.

[13] Mou W X, Celiktutan O, Gunes H. Group-level arousal and valence recognition in static images: face, body and context [C]//Proceedings of the International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Ljubljana, Slovenia: IEEE, 2015.

[14] Kosti R, Alvarez J M, Recasens A, *et al*. Emotion recognition in context [C]//Proceedings of the 30th Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017.

[15] Zhang M, Liang Y, Ma H, *et al*. Context-aware affective graph reasoning for emotion recognition [C]// Proceedings of the International Conference on Multimedia and Expo (ICME). Shanghai, China: IEEE, 2019.

[16] Bendjoudi I, Vanderhaegen F, Hamad D, *et al*. Multi-label, multi-task cnn approach for context-based emotion recognition [J]. Inform Fusion, 2020, 76: 422.

[17] Hu J, Shen L, Sun G, *et al*. Squeeze-and-excitation networks [J]. IEEE T Pattern Anal, 2018, 42: 2011.

[18] Kosti R, Alvarez J M, Recasens A, *et al*. Context based emotion recognition using emotic dataset [J]. IEEE T Pattern Anal, 2020, 42: 2755.

[19] Lin T Y, Dollar P, Girshick R, *et al*. Feature pyramid networks for object detection [C]// Proceedings of the 30th Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA: IEEE, 2017.

[20] Fukui H, Hirakawa T, Yamashita T, *et al*. Attention branch network: learning of attention mechanism for visual explanation [C]// Proceedings of the 32nd Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019.



**引用本文格式:**

中　文: 晋儒龙, 卿粼波, 文虹茜.基于注意力机制多尺度网络的自然场景情绪识别[J]. 四川大学学报: 自然科学版, 2022, 59: 012003.

英　文: Jin R L, Qing L B, Wen H Q. Emotion recognition of the natural scenes based on attention mechanism and multi-scale network [J]. J Sichuan Univ: Nat Sci Ed, 2022, 59: 012003.