Computer Science and Application 计算机科学与应用, 2022, 12(4), 1213-1225
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124123

Hans汉斯

基于标签分布学习的眼部情绪识别

李学聪，战荫伟，杨　卓，陆玉波

广东工业大学，计算机学院，广东　广州

收稿日期：2022年3月21日；录用日期：2022年4月22日；发布日期：2022年4月29日

摘　要

眼部情绪识别指的是仅从眼部区域的情绪特征识别用户的情绪状态，分析出用户佩戴智能头盔的情景下被遮挡的真实情绪。为了改善眼部区域情绪信息量少、标签歧义性所带来的识别准确度低和识别效率低的问题，本文提出一种用于识别眼部情绪的神经网络模型。该模型包含情绪标签分布生成网络和轻量级的眼部情绪识别网络两个模块。情绪标签分布生成网络会生成眼部图像的情绪分布标签，用于辅助眼部情绪识别网络的参数训练。眼部情绪识别网络包含基于注意力机制的全局特征增强模块以及局部特征增强模块，能够从信息量少的眼部图像推理出用户情绪类别。同时，为了评估网络模型性能，我们构建了2个数据集，分别是REED (Realistic Eye Emotion Datasets)和EMUG (Eye-Multimedia Understanding Group)。实验结果表明，在REED和EMUG数据集的四分类的平均准确率分别达到68.5%和80.9%，七分类的平均准确率分别达到62.0%和68.1%。同时，虽然本文模型的参数量远小于其他网络模型，但是识别效率也要优于其他模型。

关键词

眼部情绪，标签分布学习，注意力机制，特征增强，卷积神经网络

Eye Emotion Recognition Based on Label Distribution Learning

Xuecong Li, Yinwei Zhan, Zhuo Yang, Yubo Lu

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Mar. 21st, 2022; accepted: Apr. 22nd, 2022; published: Apr. 29th, 2022

Abstract

Eye emotion recognition refers to identifying the emotional state of the user only from the emotional characteristics of the eye region, and analyzing the true emotion that the user is blocked

文章引用: 李学聪, 战荫伟, 杨卓, 陆玉波. 基于标签分布学习的眼部情绪识别[J]. 计算机科学与应用, 2022, 12(4): 1213-1225. DOI: 10.12677/csa.2022.124123

when wearing the smart helmet. In order to improve the recognition accuracy and recognition efficiency caused by the lack of emotional information in the eye region and label ambiguity, a neural network model for eye emotion recognition is proposed. The model consists of two modules: emotional label distribution generation network and a lightweight eye emotion recognition network. The emotion label distribution generation network generates emotion distribution labels of eye images, which are used to assist the parameter training of the eye emotion recognition network. The eye emotion recognition network includes a global feature enhancement module and a local feature enhancement module based on the attention mechanism, which can infer user's emotion categories from eye images with less information. At the same time, in order to evaluate the performance of the network model, we constructed 2 datasets, REED (Realistic Eye Emotion Datasets) and EMUG (Eye-Multimedia Understanding Group). The experimental results show that the average accuracy of these four categories is 68.5% and 80.9% respectively, and the average accuracy of the seven categories is 62.0% and 68.1% respectively on Reed and EMUG datasets. At the same time, although the parameters of the model in this paper are much smaller than other network models, the recognition efficiency is also better than other models.

Keywords

Eye Emotion, Label Distribution Learning, Attention Mechanism, Feature Enhancement, Convolution Neural Networks

1. 引言

随着虚拟现实(VR, Virtual Reality)和眼动追踪(ET, Eye Tracking)技术的兴起，越来越多内置摄像头的智能头盔产品被推广到消费者市场中[1]。当用户佩戴智能头盔时，便能通过穿戴传感设备(数据手柄、手套)进行人机交互，享受沉浸式 VR 体验。在上述交互过程中，用户和机器分别扮演决策者和反馈者的身份，二者的关系为决策和执行，交互方式较为单一。如果能够赋予机器感知人类情绪的能力，那么便能分析出用户的使用习惯和心理状态识。通过分析用户的情绪状态，不仅能够改善智能设备系统的交互方式，而且还能指导开发厂商制定针对性的治疗方案和广告策略。因此，识别用户佩戴智能头盔下的情绪状态，对分析用户行为和改善人机交互方式具有重要意义。

当用户佩戴智能头盔时，面临两方面的挑战，一方面为面部上半区域被设备严重遮挡，难以获得有效的情绪特征；另一方面是在用户的使用过程中，产生较大的姿态变化，例如旋转和移动，导致面部特征位置定位不准确和特征提取难度较大。Ekman 等人[2]从解剖学的角度出发，表明绝大部分情绪变换都与眼部区域的肌肉运动单元有关，眼部区域蕴含着丰富的情绪信息。因此，仅利用眼部区域识别出用户的情绪状态是可行的。通过眼部区域的情绪识别，不仅能解决面部遮挡问题，提取有效的情绪特征；还能利用摄像头与眼部之间的距离和角度相对不变特性，解决姿态变化带来的问题。

眼部情绪识别方法根据是否配置额外的传感设备，分为基于传感器和基于外观的方法，前者是利用传感设备捕获眼部生理特征进行用户情绪状态识别，后者是直接通过眼部区域图像推理出用户的情绪类别。虽然基于传感器的方法能够获取更多有效的情绪特征，但是需要更改智能头盔结构并且影响用户使用体验。随着深度学习的快速发展，相关研究人员提出基于外观的方法，利用卷积神经网络解决眼部区

域的情绪识别问题。此类方法的大体思路是，首先从眼部图像中提取情绪特征，并用于预训练个性化分类器，增强模型的信息提取能力；然后，将个性化分类器加入网络框架中，并通过卷积神经网络输出情绪类别。虽然上述方法无需额外的传感器，但是以往工作提出的网络模型的参数量过于庞大且需要单独训练个性化分类器，导致模型在移动设备上运行缓慢且需要额外标定工作。

为了解决上述问题，本文提出一种基于标签分布学习的眼部情绪识别框架，主要由情绪识别网络和标签分布生成网络组成，见图 1。情绪识别网络由局部特征增强模块、全局特征增强模块和轻量级 ShuffleNet-V2 [3]骨干网络组成，用于预测眼部情绪类别；标签分布生成网络选取 ResNet50 [4]网络生成眼部图像的情绪标签分布数据集，辅助眼部情绪网络的标签分布学习。通过标签分布生成网络生成眼部图像的情绪标签分布数据，并利用标签分布数据训练轻量级眼部识别网络，从而实现轻量级网络计算效率快和识别准确度高的效果。

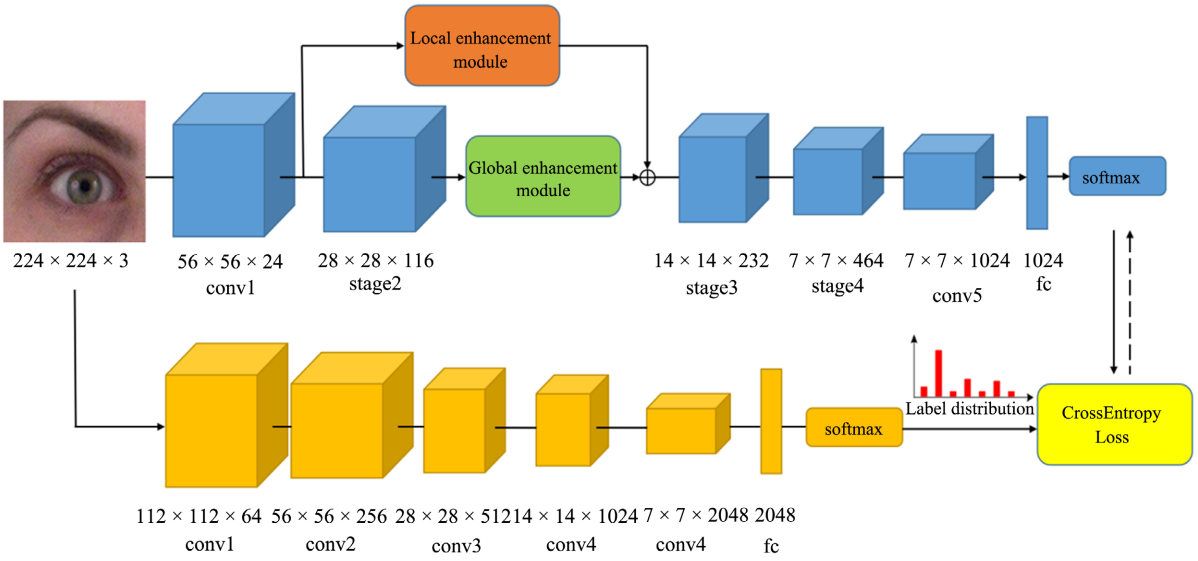

Figure 1. Eye emotion recognition network (up) and label distribution generation network (down)
图 1. 眼部情绪识别网络(上)与标签分布生成网络(下)

迄今为止，有关眼部情绪数据集尚未开源。为了评估模型的性能，本文从两个方面构建实验数据集，一方面设计和搭建眼部情绪采集方案，收集真实环境下的眼部情绪数据集 REED (Realistic Eye Emotion Datasets)；另一方面利用公开的全脸情绪数据集 MUG (Multimedia Understanding Group) [5]，从全脸中裁剪出眼部区域的图像，制作实验环境下的眼部情绪数据集 EMUG (Eye-Multimedia Understanding Group)。实验结果表明，本文方法在 REED 和 EMUG 数据集的识别准确度和识别效率均优于最先进的方法。本文主要贡献如下：

1) 提出轻量级的眼部情绪识别模型，利用注意力机制构建了局部特征强化模块和全局特征强化模块，解决轻量级网络提取信息能力不足和眼部区域情绪信息量少的问题。同时，该模型无需用户进行额外的标定工作，也能达到较优的准确度和识别效率。

2) 设计标签分布生成网络，通过自动生成眼部图像的情绪标签分布数据，辅助眼部情绪识别网络进行标签分布学习，改善数据集标签的歧义性问题，提高模型的鲁棒性。

3) 从两个方面构建实验数据集，一方面，通过设计一套眼部数据采集设备，建立真实环境下的数据集 REED；另一方面，通过公开的全脸情绪数据集 MUG，裁剪出相应的眼部情绪数据集 EMUG。实验结

果分析，本文网络模型在 REED 和 EMUG 的四分类情绪平均准确度分别为 68.5%和 80.9%，在七分类情绪平均准确度分别为 62.0%和 68.1%。

2. 相关工作

基于眼部区域的情绪识别是人脸情绪识别的一个分支，经过多年来的相关研究者的探索，取得较为丰富的研究成果。眼部情绪识别方法分为基于传感器和基于外观。

基于传感器的眼部情绪识别方法的思路是在头戴式设备配备不同的传感设备捕获用户眼部生理信号，并结合眼部图像建眼部情绪识别模型，推理出用户的情绪类别。早期的研究工作主要利用单一的眼部生理特性，并推理出眼部情绪状态。Scheirer 等人[6]采用皮肤电压传感器检测眼部肌肉的电位变化，区分用户喜悦、困惑和平静的情绪状态。Fukumoto 等人[7]利用光遮断器检测眼部肌肉的运动状态，判别用户中立、微笑和大笑的状态。Masai 等人[8] [9]采用光学传感器检测眼部肌肉运动状态，并通过 SVM [10]模型判别情绪类别。然而，单一的生理特征容易受用户自身的生理状态影响，难以准确识别用户情绪状态。随着传感设备的发展，智能头盔能够装配各式各样的传感设备采集眼部的生理特征，相关研究者提出多模态融合的眼部情绪识别方法。Kwon 等人[11]采取数据级融合的策略，首先通过传感器采集眼部的皮肤电反应和血容量搏动等生理信号数据，分别计算平均值、标准差等统计数据作为情感特征向量；接着，采用 PCA 方法[12]降低眼部图像的数据维度，提取相应的情感特征；最终，将统计特征和图像情感特征拼接融合，输入到 SVM 模型推理出情绪类别。Soleymani 等人[13]采用决策级融合的策略，将眼球注视时间、瞳孔直径和脑电图(EEG)输入到不同的分类器进行训练，并在决策阶段将分类器输出的决策特征进行融合，进而推理用户的情绪类别。Nie 等人[14]采用卷积神经网络和二叉决策树的组合策略，首先利用 AlexNet [15]提取眼部区域的眉毛内轮廓、眼睑形状以及瞳孔位置等信息，然后通过距离探测器和头部惯性传感器测量用户眼部肌肉和头部姿态的状态，最终将各种特征数据输入到二叉决策树中，进行判断用户的情绪状态。然而，上述方法存在一些不足，一方面需要配备额外的传感器，难以集成到现有的商业智能头盔；另一方面，传感器测量用户生理信号时，需要紧贴用户皮肤，影响用户的使用体验。

基于外观的眼部情绪识别方法与前者方法不同，其思路为仅从摄像头拍摄眼部区域图像识别出用户的情绪状态，无需装配额外的传感器设备。早期的工作主要利用眼球的几何特征和生理信息识别用户情绪。Babiker 等人[16]利用眼球瞳孔直径大小来区分用户的情绪状态，当用户处于中立情绪时，瞳孔直径最小，正向情绪时次之，负面时瞳孔直径最大。Nummenmaa 等人[17]利用眼球运动状态判别用户情绪状态，当用户的注视点长时间聚焦在某个区域，表明用户对该区域感兴趣。然而，上述方法使用的情绪特征变化单一，难以细粒度区分多种情绪类别。随着深度学习的发展，相关研究者利用卷积神经网络解决眼部情绪识别任务。Hickson 等人[18]利用 InceptionNet [19]提取眼部图像中眼部区域的动作信息，并根据识别出的动作组合判别用户的情绪状态。此外，该工作利用各种情绪的图像减去其中立情绪构建个性化分类器，解决不同用户情绪表达差异的问题。实验表明，该方法在自建数据集上，识别出五类情绪的平均准确率为 73.7%。Wu 等人[20]选择 ResNet18 提取眼部情绪特征，通过 Kmeans [21]聚类算法为不同用户制定个性化分类器。同时，为了提高网络识别效率，采用 SiameseNet [22]网络判断视频帧的用户是否相似，相似则跳过。最终，该方法在自建数据集上，识别七类情绪的平均准确率为 76.6%。

综上所述，基于深度学习的眼部情绪识别方法具有两点优势，一方面无需装配额外的传感器设备，另一方面能够区分更加细粒度的情绪状态。但是上述方法存在一些不足：1) 网络模型过于复杂，难以在移动设备上高效运行；2) 需要用户进行额外标定工作，并单独训练个性化分类器。因此，本文提出一种轻量级基于标签分布学习的眼部情绪识别框架，旨在无需用户进行额外标定的工作，也能达到较优的准确度和识别效率。

3. 方法

本文提出一种基于标签分布学习的眼部情绪识别框架，框架分为情绪识别网络和标签分布生成网络。情绪识别网络由局部特征增强模块、全局特征增强模块和轻量级 ShuffleNet-V2 骨干网络组成，用于预测眼部情绪类别；标签分布生成网络选取 ResNet50 网络自动生成眼部图像的情绪标签分布数据集，辅助眼部情绪网络的标签分布学习。

3.1. 眼部情绪识别网络

基于外观的眼部情绪识别方法，主要存在智能头盔计算资源有限和眼部区域情绪信息量少等问题。Hickson 等人[18]采取深度网络和个性化分类器的策略解决情绪信息量少的问题，但是需要额外标定工作和消耗大量的计算资源。为了解决上述问题，本文的眼部情绪识别网络采用 ShuffleNet-V2 轻量级网络作为主干网络，加入局部特征增强模块和全局特征增强模块，从而实现高效、准确且无需额外标定的效果。

眼部情绪识别网络首先将眼部图像初步提取情绪特征，通过局部特征增强模块和全局特征增强模块赋予有效的情绪信息特征更大的权重。然后，将局部信息和全局信息进行特征融合，并输出到后续的 Stage3、Stage4、Conv5 网络层中；最后，通过全连接层和 Softmax 层输出对应情绪类别。

3.2. 局部特征增强模块

Ekman 等人[2]从解剖学的角度出发，表明不同面部运动单元(Action Unit，AU)对应不同情绪类别。为了情绪识别网络聚焦于有效的眼部情绪 AU，本文构建基于注意力机制的局部特征增强模块，见图 2。

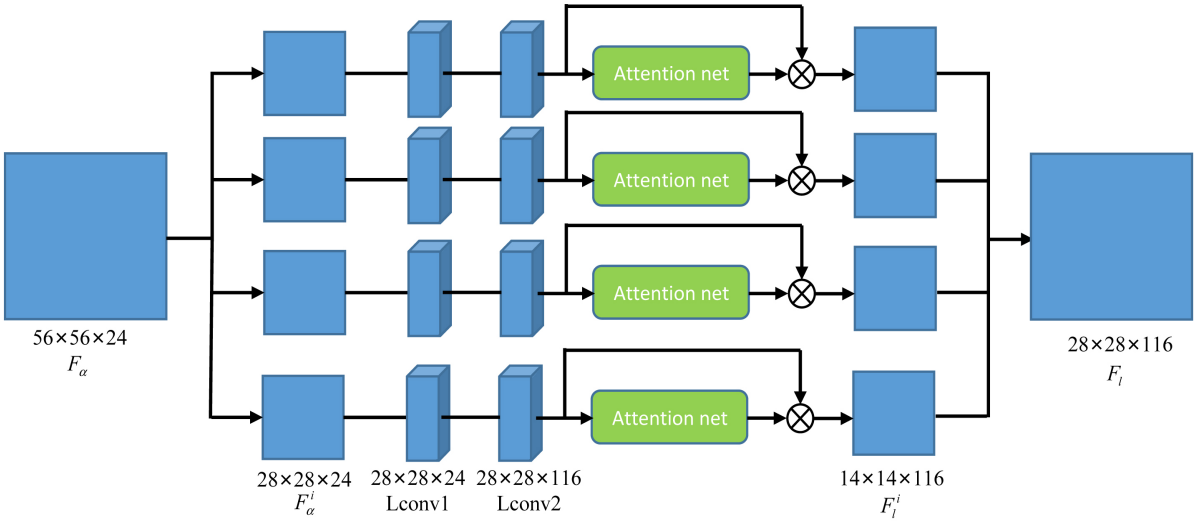

Figure 2. Local enhancement module
图 2. 局部增强模块

首先，将大小 224 × 224 × 3 的眼部区域图像输入眼部情绪识别网络中，通过 Conv1 卷积层输出大小 56 × 56 × 24 情绪特征图；然后，将情绪特征图平分为 4 个 28 × 28 × 24 的情绪特征子图，并输入到 2 个卷积层获得大小为 14 × 14 × 116 深层的情绪特征子图，其中；最后，利用 AttentionNet [23]根据 AU 重要性赋予不同特征子图的相应情绪特征权重，并与深层的情绪特征子图相乘，合并输出 28 × 28 × 116 局部增强特征图。

3.3. 全局特征增强模块

虽然局部特征增强模块增强关键 AU 的表达能力，但是缺乏捕获眼部区域 AU 之间的相关性和全局性信息的能力，不利于推断困难样本。为了解决上述问题，本文基于 Park 等人[24]提出的 BAM (Bottleneck Attention Module)，设计出全局特征增强模块，见图 3。全局特征增强模块由通道模块(Channel Module)和空间模块(Spatial Module)组成，通过提取特征图的通道信息和空间信息，寻找特征图 AU 之间的相关性。

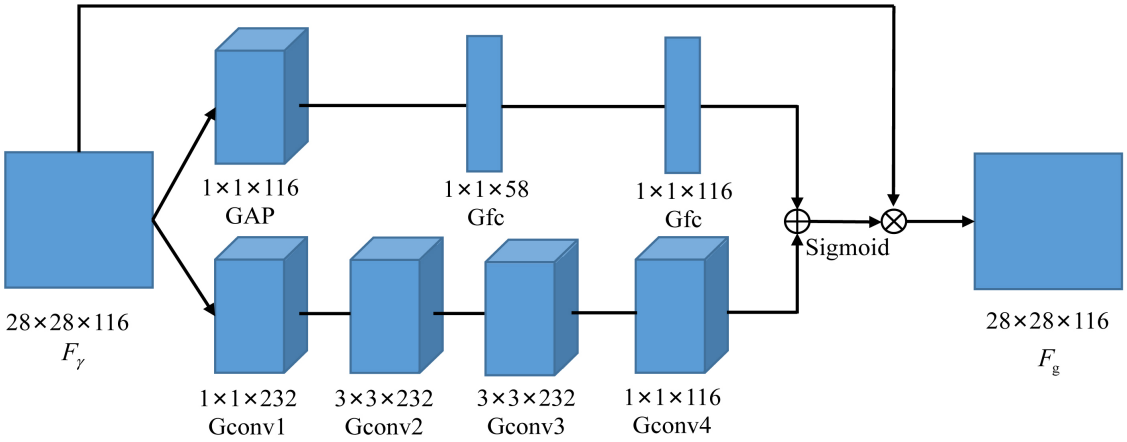

Figure 3. Channel module (up) and spatial module (down)
图 3. 通道模块(上)和空间模块(下)

特征图通过 Stage2 模块输出大小 28 × 28 × 116 的特征图，分别输入到通道模块和空间模块。通道模块的作用是关注不同通道重要性，根据任务需求进行强化或者抑制不同通道。通道模块的提取步骤是，将输入全局平均池化层 GAP 进行通道融合，并利用 2 个全连接层 Gfc 计算出大小 1 × 1 × 116 通道权重特征。空间模块是为了聚焦眼部区域 AU 的空间位置信息，计算整个特征图的空间注意力权重，强化或者抑制不同空间位置特征表达能力。空间模块基于 He 等人[4]提出瓶颈结构(Bottleneck Structure)，由 2 个 1 × 1 卷积层和 2 个 3 × 3 卷积层组成，用于提取眼部区域 AU 的空间信息。空间模块的特征提取思路是，首先利用 1 × 1 卷积层进行升维操作提高通道数量，并融合通道信息；然后，通过 2 个 3 × 3 卷积层进行空间信息变换，关联上下文信息；最后，经过 1 × 1 卷积层进行降维操作，输出大小 28 × 28 × 1 的空间权重特征。

在上述提取过程中，首先将通道模块和空间模块生成的特征权重和映射到同一维度空间，进行特征权重融合；然后，通过 Sigmoid 激活函数输出 0 到 1 之间的大小为 28 × 28 × 116 全局特征信息权重；最终，通过乘法运算将加权到每个通道的特征上，输出大小 28 × 28 × 116 的全局特征图。

至此，将眼部情绪识别网络生成的局部特征图和全局特征图进行元素求和，输入到后续 Stage3，stage4 和 Conv5 网络层中，并通过全连接层的信息融合和 Softmax 函数的概率化，输出用户情绪类别。

3.4. 情绪标签分布生成网络

标签分布学习的关键问题是如何构建标签分布数据集，通常的做法是根据样本的纹理特征和结构信息构建样本标签映射函数，从而生成相应的标签分布数据集。但上述方法仅适用特定的场景和数据，限制模型的泛化能力。Zhao 等人[25]利用卷积神经网络强大的提取能力，自动生成人脸情绪标签数据集，有效解决标注者标注错误的主观问题和图像模糊带来错误的客观问题。

受到该工作的启发，眼部情绪标签生成网络利用 ResNet50 深层网络自动生成眼部图像的情绪分布标签，辅助眼部情绪识别网络的参数训练。眼部情绪标签生成网络采用迁移学习方式，解决眼部数据的情绪信息不足和缺乏大型的公开数据集等问题，提高情绪数据标签分布的准确性。在迁移学习过程中，首先将情绪标签生成网络在公开全脸数据集 FER2013 [26]进行预训练参数权重；然后，将模型在 EMUG 数据集上进行参数微调，使其适应目标场景。

当迁移学习完成时，眼部情绪标签生成网络能够产生眼部图像的情绪标签分布，用于辅助训练眼部情绪识别网络。具体而言，给定一个图像 f 和其对应的情绪标签 $l \in \{0,1,\cdots,c-1\}$，其中 c 是情绪种类数目。首先利用情绪标签生成网络的卷积层对图像 f 进行特征提取，然后将情绪特征输入到全连接层进行特征融合，输出各类别的情绪标签向量 $v=(v_0,v_1,\cdots,v_{c-1})$；最后，利用 Softmax 函数生成情绪标签分布 $d=(d_0,d_1,\cdots,d_{c-1})$，

$$d_n=\frac{\exp(v_n)}{\sum_{m=0}^{c-1}\exp(v_m)} \tag{1}$$

其中 $n \in \{0,1,\cdots,c-1\}$。

3.5. 损失函数

根据上文所述，情绪标签生成网络的输出是真实标签分布，眼部情绪识别网络的输出是预估标签分布。因此，本文网络的损失函数采用交叉熵函数，计算真实标签分布和预估标签分布的差距，利用反向传播对眼部情绪识别网络的参数进行更新，即

$$L=-\frac{1}{N\times c}\sum_{p=0}^{N-1}\sum_{q=0}^{c-1}d_q^p\log\left(\tilde{d}_q^p\right) \tag{2}$$

其中，N 为样本数目，$\tilde{d}_q^p$ 为眼部情绪识别网络的预估情绪标签分布，上标 p 为样本序号，下标 q 为情绪类别。

4. 实验与结果分析

4.1. 实验数据集

迄今为止，有关眼部情绪数据集尚未开源。为了评估模型的性能，本文从两方面构建实验数据集，一方面是设计和搭建眼部情绪采集方案，收集真实环境下的眼部情绪数据集 REED (Realistic Eye Emotion Datasets)；另一方面是利用公开的全脸情绪数据集 MUG，通过裁剪眼部区域的图像，制作实验环境下的眼部情绪数据集 EMUG (Eye-MUG)。

4.1.1. 自建数据集

为了采集真实环境下的眼部区域图像，我们搭建一套简易的采集设备，设备主要由头戴设备、支撑力臂和摄像头组成，见图 4。通过头盔和可调节支撑力臂的组合，能够自由调节志愿者的眼部区域与摄像头间的角度和距离，便于数据采集。同时，为了减少外界光照影响和提高图像质量，选择带 6 个发光二极管的红外且分辨率为 640 × 480 的摄像头，其中采集频率为 30 帧/秒。

REED 的数据采集工作共征集 25 位志愿者，其中男性 14 名、女性 11 名。在数据采集的过程中，首先要求志愿者佩戴采集设备观看情绪诱导图片，激发和指导志愿者做出相应的情绪状态；然后，利用 OpenCV 图像处理工具收集志愿者的眼部区域视频并标注相应的情绪类别；最终，清理闭眼和情绪切换时产生的错误情绪图像数据。

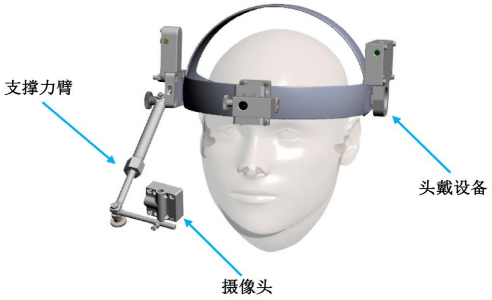

Figure 4. Schematic diagram of eye data acquisition equipment
图 4. 眼部数据采集设备示意图

每位志愿者采集 7 类情绪，分别为悲伤、恐惧、惊讶、愤怒、快乐、厌恶、平静，见图 5。每类情绪需要采集 2 次，每次采集拍摄 7 秒时长的视频，数据采集总时长为 2450 秒，图像总数为 73,500 张。通过数据清理，筛选出 53,375 张图像，其中每位志愿者的每类情绪图像约为 305 张。对于筛选出来的图像，根据所属志愿者的不同进行分组，从中随机挑选 10 组志愿者的数据作为训练集，5 组志愿者的数据作为验证集，10 组志愿者的数据作为测试集。

Figure 5. REED dataset, seven eye emotion images from the same volunteer
图 5. REED 数据集，同一志愿者的七种眼部情绪图像

4.1.2. 裁剪数据集

从全脸情绪数据集裁剪出眼部情绪数据集的要求有两方面，一方面需要图像分辨率高和无遮挡的图像，才能裁剪出高质量的眼部区域图像；另一方面需要稳定的光照条件和头部姿态，才能模拟佩戴智能头盔下的场景。因此，本文选取全脸情绪数据集 MUG 制作眼部情绪数据集 EMUG。

MUG 公开数据集共有 52 位志愿者的全脸情绪图像，情绪类别共 7 类，分别是悲伤、恐惧、惊讶、愤怒、快乐、厌恶、平静，见图 6。每类情绪收集 3 到 5 组的视频序列，每组序列往往有 50 到 150 张图像。制作 EMUG 的过程中，利用 Eivazi 等人[27]提出的眼部区域检测方法将 MUG 的全脸情绪图像裁剪出相应的眼部区域，使用手工剔除闭眼和错误情绪的图像，并将每张图像打上相应的情绪类别标签。EMUG 数据集共有 52 位志愿者，包含七类情绪，分别是悲伤、恐惧、惊讶、愤怒、快乐、厌恶、平静，见图 7。其中，每位志愿者的每类情绪图像约为 184 张，数据集总共包含 66,976 张图像。为了更加合理测试网络性能，EMUG 数据集根据所属志愿者的不同进行分组，从中随机挑选 28 组志愿者的数据作为训练集，12 组志愿者的数据作为验证集，12 组志愿者的数据作为测试集。

Figure 6. Seven facial emotions from the same volunteer from the public data set MUG
图 6. 来自公开数据集 MUG 同一志愿者的七种面部情绪

Figure 7. Expose the EMUG dataset image of the culled dataset MUG
图 7. 公开数据集 MUG 经过裁剪后的 EMUG 数据集图像

4.2. 实验设置

本文对所有数据集预处理操作，将图像大小调整为 224 × 224 像素，便于模型的训练和评估，利用随机裁剪和水平翻转等手段，增加样本的多样性和减少模型的过拟合。实验模型的损失函数选择带动量优化的随机梯度下降算法(Stochastic Gradient Descent, SGD)，并将动量值设置为 0.8。同时，模型初始学习率设置为 0.0001，训练批次样本数量为 128，共需要 600 次周期训练。

实验的硬件配置如下，CPU 为 AMD Ryzen 5 3600、GPU 为 NVIDIA GeForce RTX2070s 8GB、RAM 为 32GB，并采用 Pytorch 1.9 搭建、训练和评估网络模型。

4.3. 评估指标

本文模型的评估指标采用精确率 P 和 $F1$ 评分。准确率 P，召回率 R 和 $F1$ 评分的计算公式为：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$F1 = \frac{2PR}{P + R} \tag{5}$$

其中，TP 是真正例，FP 是假正例，FN 是假负例。

本文的网络模型学习方式分为单标签学习和多标签学习，单标签学习的模型输出预测标签是单元组；虽然多标签学习模型输出的预测标签是多元组，但会采用最大值函数输出多元组里的最大概率值，即输出预测值仍为单标签。

4.4. 对比模型介绍

为了验证本文方法的有效性，本文对比相近工作的 Eyemotion [18]和 EMO [20]模型。

Eyemotion 采用其他情绪类别减去平静情绪类别的像素值的策略构建个性化分类器，增强提取眼部区域的情绪特征，并将个性化分类器加入 Inception 网络识别用户情绪。该方法支持 5 类情绪识别，分别是愤怒、惊讶、快乐、平静和眯眼。由于眯眼更多体现眼部生理上的运动状态，难以体现用户的情绪状态，因此不在实验的讨论范围内。

EMO 采用 Kmeans 聚类方法构建个性化分类器，并将个性化分类器加入 ResNet18 网络输出情绪状态。该方法支持 7 类情绪识别，分别是悲伤、恐惧、惊讶、愤怒、快乐、厌恶、平静。

迄今为止，Eyemotion 和 EMO 是最接近本文工作的内容——基于眼部区域的情绪识别，但是二者所使用的模型、训练代码和实验数据尚未公开。因此，本文复现了二者的网络模型。因为 Eyemotion 和 EMO 需要在微调阶段前构建个性化分类器模块，所以提前进行额外的标定工作。本文方法和前者两种方法的训练策略都采用迁移学习方法，在预训练阶段，三种方法均选用 ImageNet 数据集进行参数更新。在微调

阶段，利用 REED 数据集和 EMUG 数据集的训练集和验证集对三种方法的网络模型进行参数微调，并使用测试集对模型进行性能评估。

4.5. 实验结果与分析

4.5.1. 对比实验

三种方法分别在 REED 和 EMUG 数据集上，进行四分类情绪和七分类情绪准确度对比实验。

四分类情绪识别在 REED 和 EMUG 数据集上的实验结果，见表 1。结果表明，本文方法的 F1 值分别达到 0.720 和 0.826，准确率分别达到 68.5%和 80.9%。相比较于最优模型 EMO，本文模型的 F1 指标分别提高了 0.067 和 0.042，平均准确度分别提高了 3.8%和 4.0%。其中，三种方法在 REED 数据集的准确度要明显低于 EMUG 数据集，原因有两方面，一方面是 REED 数据集样本数目要少于 EMUG 数据集，另一方面是实际环境数据集存在标签歧义性和图像质量不佳的问题。

Table 1. Experimental results of four categories of eye emotion recognition
表 1. 四分类眼部情绪识别实验结果

数据集	方法	F1 分数	准确率(%)
REED-4	EMO	0.653	64.7
REED-4	Eyemotion	0.601	57.9
REED-4	**Ours**	**0.720**	**68.5**
EMUG-4	EMO	0.784	76.9
EMUG-4	Eyemotion	0.703	66.8
EMUG-4	**Ours**	**0.826**	**80.9**

七分类情绪识别在 REED 和 EMUG 数据集的实验结果，见表 2。结果表明，本文方法的 F1 值分别达到 0.658 和 0.702，准确率分别达到 62.0%和 68.1%。相比较于最优模型 EMO，本文模型的性能表现更佳。其中，七分类情绪的准确度低于四分类的准确度，原因是存在部分情绪类别的眼部 AU 表达类似，例如恐惧和惊讶、愤怒和厌恶，使样本标签具有歧义性，增加识别难度。

综上所述，本文所提的方法模型优于其他方法的原因是，一方面标签分布学习减少了数据集带来的数据标签歧义性影响；另一方面局部模块和空间模块具有更强的特征提取能力，能够在实际环境中取得更佳的性能。

Table 2. Experimental results of seven categories of eye emotion recognition
表 2. 七分类眼部情绪识别实验结果

数据集	方法	F1 分数	准确率(%)
REED-4	EMO	0.653	64.7
REED-4	Eye motion	0.601	57.9
REED-4	**Ours**	**0.720**	**68.5**
EMUG-4	EMO	0.784	76.9
EMUG-4	Eye motion	0.703	66.8
EMUG-4	**Ours**	**0.826**	**80.9**

4.5.2. 消融实验

为了验证各个组件的有效性，将在 REED 数据集和 EMUG 数据集上进行四分类情绪识别的组件消融实验，结果见表 3。消融实验选择 ShuffleNet-V2 网络模型作为基线网络，分别将局部增强模块(Local Enhancement Module, LEM)和全局增强模块(Global Enhancement Module, GEM)添加到基线网络中，进行组合测试。当基线网络仅添加局部特征增强模块时，在 REED 数据集和 EMUG 数据集上的四分类情绪识别准确度分别比基线网络提高了 6.6%和 8.2%。当基线网络仅添加全局增强模块时，在 REED 数据集和 EMUG 数据集上的四分类情绪识别准确度分别比基线网络提高了 6.1%和 7.7%。当基线网络情加入了局部特征增强模块和全局特征增强模块，即本文所提出的眼部情绪识别网络，虽然网络模型增加了微小的计算开销，但性能提升明显。

Table 3. Experimental results of different modules
表 3. 不同模块实验结果

数据集	方法	参数量(M)	MFLOPs	准确率(%)
REED-4	Baseline	1.26	147.79	56.8
REED-4	Baseline + LEM	1.27	152.79	63.4
REED-4	Baseline + GEM	1.27	153.79	62.9
REED-4	**Baseline + LEM + GEM**	**1.28**	**154.18**	**66.7**
EMUG-4	Baseline	1.26	147.79	65.2
EMUG-4	Baseline + LEM	1.27	152.79	73.4
EMUG-4	Baseline + GEM	1.27	153.79	72.9
EMUG-4	**Baseline + LEM + GEM**	**1.28**	**154.18**	**77.1**

为了验证标签分布学习方法的有效性，选择在 REED 数据集和 EMUG 数据集进行不同学习方式的四分类情绪识别消融实验，比较基于标签分布生成网络(Label Distribution Generator Network, LDGN)生成标签分布概率的标签分布学习(LDL)与传统的单标签学习(SLL)的性能优劣，结果见表 4。消融实验的基线网络选择 EMO 模型。本文的方法在 REED 数据集和 EMUG 数据集的结果表明，标签分布学习和单标签学习的准确度分别比基线网络提高了 1.8%和 3.8%。同时，基线网络通过标签分布学习方法，网络性能也有一定的提升，在 REED 数据集和 EMUG 数据集的准确度分别提高了 3.1%和 1.6%。值得注意的是，基于标签分布学习的情绪识别网络的参数量远低于 LDG 的参数量，但在性能上优于 LDG。基于标签分布学习的方法实现高准确度的原因有两方面，一方面是人类情绪是由多个基本情绪组合而成，标签分布更加符合客观世界；另一方面，LDG 生成的标签分布能够减少标签歧义性和人为标注错误。

Table 4. Experimental results of different modules
表 4. 不同模块实验结果

数据集	方法	参数量(M)	MFLOPs	准确率(%)
REED-4	LDGN (SLL)	23.52	4109.48	67.4
REED-4	Baseline (SLL)	11.69	2054.52	64.7
REED-4	Baseline (LDL)	11.69	2054.52	67.8
REED-4	ours (SLL)	1.28	154.18	66.7

Continued

REED-4	**ours (LDL)**	**1.28**	**154.18**	**68.5**
EMUG-4	LDGN (SLL)	23.52	4109.48	78.0
EMUG-4	Baseline (SLL)	11.69	2054.52	76.9
EMUG-4	Baseline (LDL)	11.69	2054.52	78.5
EMUG-4	ours (SLL)	1.28	154.18	77.1
EMUG-4	**ours (LDL)**	**1.28**	**154.18**	**80.9**

5. 结论

为了解决眼部情绪识别信息量少和识别效率慢的问题，本文提出一个基于标签分布学习的眼部情绪识别框架。网络框架由眼部情绪识别网络和情绪标签生成网络组成，通过情绪标签生成网络输出眼部图像的情绪标签分布数据，对轻量级的眼部情绪识别网络进行参数更新，使得轻量级的眼部情绪识别网络实现较优的准确度和识别效率。同时，引入基于注意力机制的局部特征强化模块和全局特征强化模块，增强眼部情绪识别网络的特征提取能力，进一步网络的识别准确度。此外，为了评估网络的性能，构建了 REED 和 EMUG 眼部情绪数据集。实验结果表明，本文模型在 REED 和 EMUG 数据集上的识别准确度和识别效率均优于同类别的方法。

基金项目

广东省重点领域研发计划项目(编号 2020B0101130019，2019B010150002)，国家自然科学基金(批准号 61907009)。

参考文献

[1] Lorenz, O. and Thomas, U. (2019) Real Time Eye Gaze Tracking System Using CNN-Based Facial Features for Human Attention Measurement. *Proceedings of the* 14*th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Vol. 5, Prague, 25-27 February 2019, 598-606. https://doi.org/10.5220/0007565305980606

[2] Friesen, W.V. and Ekman, P. (1983) EMFACS-7: Emotional Facial Action Coding System. Unpublished Manuscript, University of California at San Francisco, San Francisco.

[3] Ma, N., Zhang, X., Zheng, H.T. and Sun, J. (2018) Shufflenet v2: Practical Guidelines for Efficient CNN Architecture Design. *Proceedings of the European Conference on Computer Vision*, Munich, 8-14 September 2018, 122-138. https://doi.org/10.1007/978-3-030-01264-9_8

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

[5] Aifanti, N., Papachristou, C. and Delopoulos, A. (2010) The MUG Facial Expression Database. 11*th International Workshop on Image Analysis for Multimedia Interactive Services* (*WIAMIS* 10), Desenzano del Garda, 12-14 April 2010, 1-4.

[6] Scheirer, J., Fernandez, R. and Picard, R.W. (1999) Expression Glasses: A Wearable Device for Facial Expression Recognition. *CHI*'99 *Extended Abstracts on Human Factors in Computing Systems*, Pittsburgh, 15-20 May 1999, 262-263. https://doi.org/10.1145/632716.632878

[7] Fukumoto, K., Terada, T. and Tsukamoto, M. (2013) A Smile/Laughter Recognition Mechanism for Smile-Based Life Logging. *Proceedings of the* 4*th Augmented Human International Conference*, Stuttgart, 7-8 March 2013, 213-220. https://doi.org/10.1145/2459236.2459273

[8] Masai, K., Kunze, K., Sugimoto, M. and Billinghurst, M. (2016) Empathy Glasses. *Proceedings of the* 2016 *CHI Conference Extended Abstracts on Human Factors in Computing Systems*, San Jose, 7-12 May 2016, 1257-1263. https://doi.org/10.1145/2851581.2892370

[9] Masai, K., Sugiura, Y., Suzuki, K., Shimamura, S., Kunze, K., Ogata, M., *et al*. (2015) AffectiveWear: Towards Recognizing Affect in Real Life. *Adjunct Proceedings of the* 2015 *ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the* 2015 *ACM International Symposium on Wearable Computers*, Osaka, 7-11 September 2015, 357-360. https://doi.org/10.1145/2800835.2800898

[10] Cherkassky, V. and Ma, Y. (2004) Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. *Neural Networks*, **17**, 113-126. https://doi.org/10.1016/S0893-6080(03)00169-2

[11] Kwon, J., Ha, J., Kim, D.H., Choi, J.W. and Kim, L. (2021) Emotion Recognition Using a Glasses-Type Wearable Device via Multi-Channel Facial Responses. *IEEE Access*, **9**, 146392-146403. https://doi.org/10.1109/ACCESS.2021.3121543

[12] Yang, J., Zhang, D., Frangi, A.F. and Yang, J.-Y. (2004) Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 131-137. https://doi.org/10.1109/TPAMI.2004.1261097

[13] Soleymani, M., Pantic, M. and Pun, T. (2011) Multimodal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing*, **3**, 211-223. https://doi.org/10.1109/T-AFFC.2011.37

[14] Nie, J., Hu, Y., Wang, Y., Xia, S. and Jiang, X. (2020) SPIDERS: Low-Cost Wireless Glasses for Continuous *In-Situ* Bio-Signal Acquisition and Emotion Recognition. 2020 *IEEE/ACM* 5*th International Conference on Internet-of-Things Design and Implementation* (*IoTDI*), Sydney, 21-24 April 2020, 27-39. https://doi.org/10.1109/IoTDI49375.2020.00011

[15] Yuan, Z.W. and Zhang, J. (2016) Feature Extraction and Image Retrieval Based on AlexNet. *Eighth International Conference on Digital Image Processing* (*ICDIP* 2016), International Society for Optics and Photonics, 100330E.

[16] Babiker, A., Faye, I., Prehn, K. and Malik, A. (2015) Machine Learning to Differentiate between Positive and Negative Emotions Using Pupil Diameter. *Frontiers in Psychology*, **6**, Article No. 1921. https://doi.org/10.3389/fpsyg.2015.01921

[17] Nummenmaa, L., Hyönä, J. and Calvo, M.G. (2006) Eye Movement Assessment of Selective Attentional Capture by Emotional Pictures. *Emotion*, **6**, 257-268. https://doi.org/10.1037/1528-3542.6.2.257

[18] Hickson, S., Dufour, N., Sud, A., Kwatra, V. and Essa, I. (2019) Eyemotion: Classifying Facial Expressions in VR Using Eye-Tracking Cameras. 2019 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 7-11 January 2019, 1626-1635. https://doi.org/10.1109/WACV.2019.00178

[19] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[20] Wu, H., Feng, J., Tian, X., Sun, E., Liu, Y., Dong, B., *et al*. (2020) EMO: Real-Time Emotion Recognition from Single-Eye Images for Resource-Constrained Eyewear Devices. *Proceedings of the* 18*th International Conference on Mobile Systems, Applications, and Services*, Toronto, 15-19 June 2020, 448-461. https://doi.org/10.1145/3386901.3388917

[21] Krishna, K. and Murty, M.N. (1999) Genetic K-Means Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* (*Cybernetics*), **29**, 433-439. https://doi.org/10.1109/3477.764879

[22] Tang, S., Andriluka, M., Andres, B. and Schiele, B. (2017) Multiple People Tracking by Lifted Multicut and Person re-Identification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 3701-3710. https://doi.org/10.1109/CVPR.2017.394

[23] Yoo, D., Park, S., Lee, J.Y., Paek, A.S. and Kweon, I.S. (2015) Attentionnet: Aggregating Weak Directions for Accurate Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 2659-2667. https://doi.org/10.1109/ICCV.2015.305

[24] Park, J., Woo, S., Lee, J.Y., *et al*. (2018) Bam: Bottleneck Attention Module. arXiv preprint arXiv:1807.06514,.

[25] Zhao, Z., Liu, Q. and Zhou, F. (2021) Robust Lightweight Facial Expression Recognition Network with Label Distribution Training. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 3510-3519. https://doi.org/10.1007/978-3-642-42051-1_16

[26] Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., *et al*. (2013) Challenges in Representation Learning: A Report on Three Machine Learning Contests. *International Conference on Neural Information Processing*, Daegu, 3-7 November 2013, 117-124. https://doi.org/10.1007/978-3-642-42051-1_16

[27] Eivazi, S., Santini, T., Keshavarzi, A., Kübler, T. and Mazzei, A. (2019) Improving Real-Time CNN-Based Pupil Detection through Domain-Specific Data Augmentation. *Proceedings of the* 11*th ACM Symposium on Eye Tracking Research & Applications*, Denver, 25-28 June 2019, Article No. 40. https://doi.org/10.1145/3314111.3319914