[文章编号] 1007－7405(2025)02－0186－12 DOI:10.19715/j.jmuzr.2025.02.11

基于动态卷积的标签不确定性自学习预测分配算法的面部表情识别

杨远奇[1]，蔡岱立[2]，谢泽凌[1,3]，江恩杰[1]

（1. 集美大学诚毅学院 信息工程系，福建 厦门 361021；2. 深圳市浩瀚卓越科技有限公司，广东 深圳 518071；
3. 自然资源部宁德海洋中心，福建 宁德 352100）

[摘要] 为解决在表情识别领域中数据集受到噪声、标注模糊、微表情等不确定性因素干扰的问题，提出了一种标签不确定性自学习预测分配算法。该算法包含三个核心模块：1）自注意力加权模块采用动态卷积实现精细的像素级注意力机制，有效降低计算负担；2）正则化排序模块通过样本权重排序和重新分配，优化了模型对不确定样本的处理；3）标签再分配模块对低权重样本进行标签校正，提高了整体预测精度。经过实验验证，该算法有效抑制了标签不确定性的影响，在RAF-DB和MMAFEDB等公开数据集上展现了卓越性能。

[关键词] 面部表情识别；标签不确定性自学习预测分配算法；动态卷积；抗噪神经网络

[中图分类号] TP 391.4

Facial Expression Recognition Based on Self-Learning Label Prediction and Distribution Algorithm Based on Dynamic Convolutional for Label Uncertainty

YANG Yuanqi[2], CAI Daili[1], XIE Zeling[1,3], JIANG Enjie[1]

(1. Information Engineering, Chengyi College, Jimei University, Xiamen 361021, China;
2. Shenzhen Hohem Technology Co., Ltd., Shenzhen 518071, China;
3. Ningde Marine Center, MNR, Ningde 352100, China)

Abstract: To solve the problem of uncertain factors such as noise, fuzzy labeling, and micro expressions affecting the dataset in the field of facial expression recognition, a label uncertainty self-learning prediction allocation algorithm is proposed. The algorithm consists of three core modules: 1) A self-attention weighting module employing dynamic convolution to achieve fine-grained pixel-level attention mechanism, effectively reducing computational overhead; 2) A regularized ranking module that optimizes the model's handling of uncertain samples through sample weight reranking and reallocation; 3) A label reassignment module that corrects labels for low-weight samples, thereby improving overall prediction accuracy. Experimental validation demonstrates the algorithm's efficacy in mitigating the impact of label uncertainty, exhibiting outstanding performance on publicly available datasets such as RAF-DB and MMAFEDB.

[收稿日期] 2023－12－15

[基金项目] 中国高校产学研创新基金——新一代信息技术创新项目“基于SDN的车联网流量预测算法研究”（2021ITA06004）；福建省中青年教师教育科研项目“车联网IoV流量预测算法研究”（JAT210671）；集美大学诚毅学院中青年项目“基于transformer的推荐算法研究”（c13019）

[作者简介] 杨远奇（1982—），副教授，从事软件定义网络、智能算法和推荐系统研究。E-mail:gigkin6808@jmu.edu.cn

Keywords: facial expression recognition; label uncertainty self-learning prediction allocation algorithm; dynamic convolution; anti-noise neural network

0　引言

面部表情识别（facial expression recognition，FER）是一项跨学科研究，对于计算机解析人类意图至关重要[1]。此领域借助深度学习已获显著进展，尽管如此，识别精度与效率的均衡仍待提高。环境变量亦复杂化了计算机对面部表情的解读。FER 技术分为处理静态与动态图像两大类，其中多模态方法也得到应用。传统技术多依赖手动特征抽取或初级学习，面对复杂情境时效果受限。

目前广泛应用的面部数据集如 CK + [2]、MMI[3]及 Oulu CASIA[4]均面临标签不一致及图像质量变异导致的不确定性，从显著至微妙表情的识别难度梯度加剧了此问题。这种标签混淆不仅妨碍模型学习关键面部特征，还可能引发过度拟合、学习难度加大及优化困难。解决这些标签问题，提升模型学习面部特征的效率，是 FER 系统面临的主要挑战。

针对这一问题，本研究在解决大规模面部表情识别中的不确定性因素方面，重新设计和改进了当今流行的抗噪神经网络模型，提出了一种名为标签不确定性自学习预测分配算法（label uncertainty self-learning prediction allocation，LUSLPA）的方法。

1　标签不确定性自学习预测分配算法

本研究提出了一种新的标签不确定性自学习预测分配算法，它包含三个核心模块：自注意力加权模块、正则化排序模块和标签再分配模块。首先，自注意力加权模块通过动态卷积技术，有效地从面部特征中提取并学习样本权重。该模块利用动态卷积机制，自适应地识别不同样本的重要性，并据此为每个样本赋予相应的权重。然后，正则化排序模块被设计来进一步处理经自注意力加权得到的样本权重。通过对权重进行排序，正则化排序模块将样本集分割为低权重组和高权重组，并引入了边际损失函数 RR-Loss 以正则化注意力权重分配，确保注意力机制在优化过程中的有效性和意义性。最后，标签再分配模块从低权重组中筛选样本，并将这些样本的最大预测概率与原始标签的概率进行比较。如果最大预测概率超过设定阈值的原始标签概率，便将该样本的标签更改为具有最大预测概率的标签。此举不仅能纠正标签的潜在错误，还能进一步提升预测精度。整体而言，本算法通过融合自注意力加权、正则化排序和标签再分配的方法，有效地抑制了数据不确定性，优化了处理噪声和异常值的能力，从而显著提高了预测的准确性。如图 1 所示，该算法的架构设计确保了对标签不确定性的高效处理。

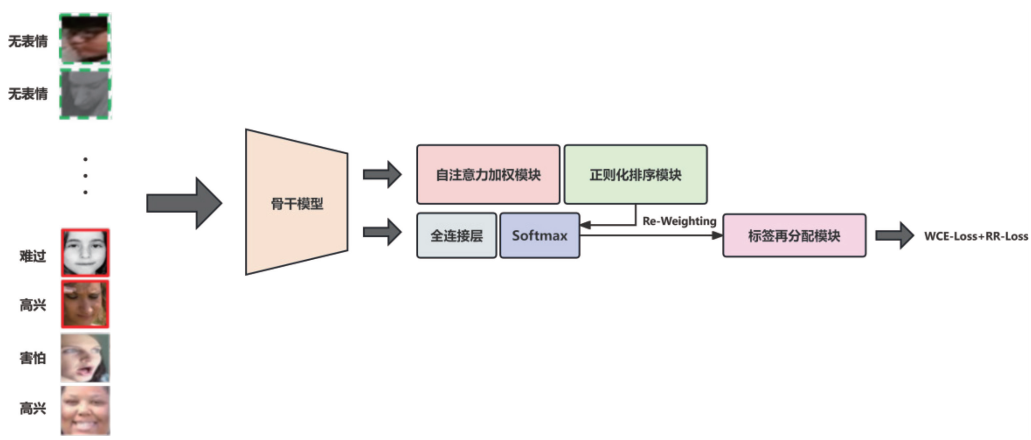

图 1　标签不确定性自学习预测分配算法的整体框架结构图

Fig.1　The overall framework structure of self-learning label prediction and distribution algorithm for label uncertainty

1.1 自注意力加权模块

本模块旨在捕获训练样本的自注意力分布，并优化特征提取与排序过程。为此，本文将样本分成两组：一组样本具有较高的权重，另一组则权重较低，以促进排序部分的顺利运行。给定一个批量样本集 $F\{x_1,x_2,\cdots,x_n\}$，通过自注意力加权模块，可以为每个样本生成注意力分布，然后将其输入正则化排序模块（见图 2）。

在自注意力特征提取方面，本文采用动态卷积网络，通过调整注意力权重以改变每个卷积核的贡献，优化传统卷积网络的特征提取方式（见图 3）。动态卷积允许注意力分布根据 F 对卷积核进行非线性加权聚合，从而提升特征提取能力。由于内核尺寸较小，聚合多个平行的卷积核既高效又具备更强的位置表达能力，同时不增加网络的深度和宽度。卷积组的聚合使用 Softmax 函数得出的结果，因此整个过程符合数据驱动特性。在整个网络训练中，学习过程自然融入其中，无需额外的空间和计算成本。

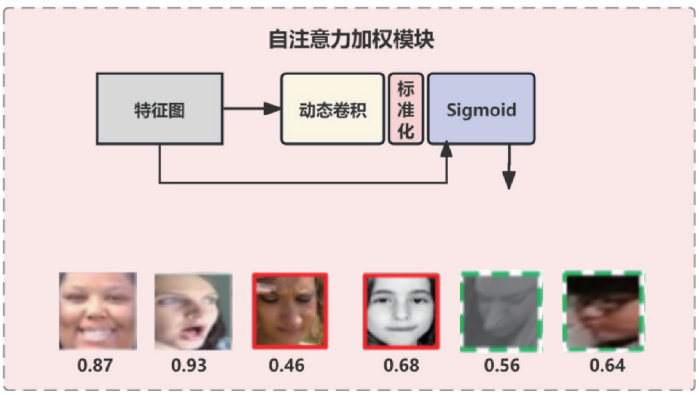

图 2 自注意力加权模块

Fig. 2 Self-attention importance weighting module

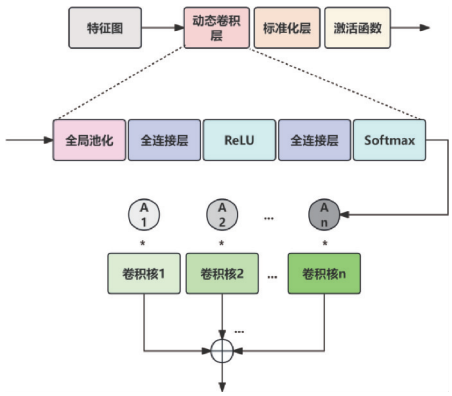

图 3 动态卷积模块结构图

Fig. 3 Structure of the dynamic convolution module

然而，动态卷积网络面临着多个卷积核和注意力模型共同学习的挑战，这个问题随着网络深度的增加变得更加突出。为了解决这个问题，本文借鉴了 Chen 等[5]的解决方案并取得了显著效果，即限制注意力的取值范围。通过限制注意力值，可以简化学习过程，减少卷积核的叠加范围。本文将注意力值限制在 0 到 1 之间，并确保所有注意力值之和为 1。为此，本文采用了带温度系数的 Softmax 函数[6]，替代了原先的 Softmax 函数。通过调整温度系数，本文可以实现类似于均匀分布的效果，即均匀分布系数在每个卷积核中。

为了缓解网络深度导致的参数和稳定性问题，本文引入了残差连接。残差连接提供了一种天然的恒等映射，允许信号跨层传递，从低级信号直接传递到高级信号，从而解决神经网络退化的问题。这种方式有效缓解了梯度相关的问题，提高了网络的稳定性和性能。即残差网络可以看作是一系列路径集合组成的集成模型，其中不同路径包含不同的网络层子集。网络性能与路径平滑系数相关，这表明残差网络的路径具有一定的独立性和冗余性。因此，残差连接可以被视为一种隐藏的集成模型。整体用公式表示为 $x = \mathrm{DyConc}(f), x = \mathrm{BN}(x) = \dfrac{x-\mu(x)}{\sqrt{\sigma^2(x)+\epsilon}} \times \gamma + \beta, x = f + \sigma(x) = f + \dfrac{1}{1+\mathrm{e}^{-x}}$。其中：$f$ 是主干模型的输出特征；DyConv 是动态卷积层；BN 是批归一化层；σ 是 sigmoid 函数，用于选择需要保留的特征。DyConv 表示为 $d_{\mathrm{w}} = \max\left(0, \dfrac{w\sum_{i=1}^{H}\sum_{j=1}^{W} x_{i,j}}{H \times W} + b\right)$；$d_{\mathrm{w}_i} = \dfrac{\mathrm{e}^{d_{\mathrm{w}_j}}}{\sum_{j=1}^{k}\mathrm{e}^{d_{\mathrm{w}_j}}}$，for $j = 1,\cdots,k$；$y = \sum_{i=1}^{C_{\mathrm{out}}}\sum_{j=1}^{C_{\mathrm{in}}}\sum_{p=1}^{K_{\mathrm{H}}}\sum_{q=1}^{K_{\mathrm{W}}} w'_{ijpq} \times x_{i+p-1,j+q-1} + b'$。其中：$w' = d_{\mathrm{w}}^{\mathrm{T}} W_{\mathrm{conv}}$；$b' = d_{\mathrm{w}}^{\mathrm{T}} b_{\mathrm{conv}}$ 是输入特征图；w 和 b 分别表示注

意力权重和注意力偏置；C_{in}和C_{out}是输入和输出通道维度的大小；K_H 和 K_w 是卷积核的尺寸；W_{conv}和b_{conv}分别是可学习的卷积核权重与偏置；d_w 是通过注意力函数计算得到的卷积核注意力向量。本文将以上整个模块作为一个 Block，利用不同的卷积核大小 k 和空洞率 d 以提取不同粗细粒度的尺度特征和聚合，用公式表示为 $y_1 = \text{Block}(x,k_1,d_1), y_2 = \text{Block}(x,k_2,d_2), y_3 = \text{Block}(x,k_3,d_3), y = \sigma(W \times [y_1,y_2,y_3] + b)$。其中：$d_1$ ，d_2 和、d_3 分别表示用不同的空洞率的动态卷积模块的结果；[ ] 为聚合操作；σ 是非线性激活函数。

1.2　正则化排序模块

本文引入了一个正则化排序模块，旨在对学习到的注意力权重进行有序化处理（见图 4）。在该模块内部机制中，首先将学习到的自注意力权重按照降序排列，并根据预设的比例参数 β 将其精确地分为高权重组和低权重组。考虑到同一表情可能由不同的面部区域引发，采用了直接约束自注意力权重的策略。另外，尽管在正则化排序模块中引入注意力机制是一种可选项，但这会增加额外的计算和参数成本。鉴于自注意力加权模块已经足够涵盖了整体样本的信息，即使在正则化排序模块中添加额外的注意力机制，也不会带来明显的增益。公式表示为 $\alpha_H = (\alpha_i \mid \alpha_i > t), \alpha_L = (\alpha_i \mid \alpha_i \leqslant t), \bar{\alpha}_H = \frac{1}{|\alpha_H|}\sum_{i\in\alpha_H}\alpha_i, \bar{\alpha}_L = \frac{1}{|\alpha_L|}\sum_{i\in\alpha_L}\alpha_i$。其中：$t$ 是权重分组的阈值。

图 4　正则化排序模块

Fig. 4　Rank regularization module

1.3　标签再分配模块

在正则化排序模块中，每个批次的样本划分为两个组：高权重组和低权重组。通过深入的实验观察，发现那些具有不确定性的样本往往呈现出较低的重要性和权重，相反，确定性样本则表现出更高的注意力值。本文思考一个直观的解决方案：设计一种方法来重新分配低权重组样本的标签。

同时引入了标签再分配模块。这个模块专门针对那些不确定性样本，并比较它们的最大预测概率与原始标签的概率。如果某个样本的最大预测概率超过了设定的阈值，那么就会将该样本的预测结果调整为模型认为最有可能的类别标签（见图 5）。换句话说，新的标签将是模型对该样本预测的最有信心的结果。这个操作的目的是确保对低权重组样本标签的重新分配是基于模型高置信度的预测结果，从而提升模型的准确性和可信度。因此，标签再分配模块的主要目的是根据概率阈值对低权重组样本进行标签的重新分配，以确保标签的准确性和模型的可靠性。整体用公式表示为 $p = \text{softmax}(y)$；$p_{max}, \hat{y} = \arg\max_c p_c$；$p_{gt} = p_t$；$m = (p_{max} - p_{gt} > m_2$；$I_i = \text{Index}(\text{nonzero}(m))$；$L = \text{update labels}(\hat{y}_i, i)$。其中：$p$ 是模型预测的概率分布；p_{max}和 $\hat{y}$ 分别是最大概率及其类别；p_{gt}是真实类别的预测概率；m 是判断是否需要更新标签的掩码；I_i 是需要更新的样本索引；L 表示标签被更新之后的数据集。

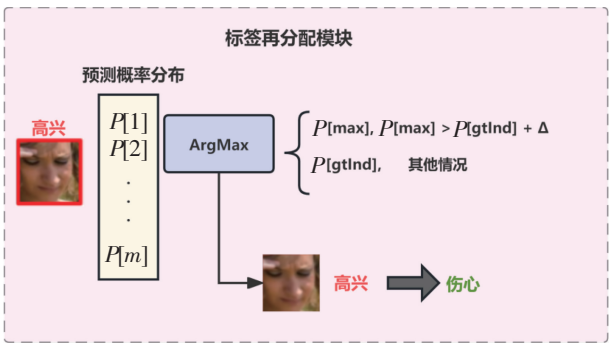

图 5　标签再分配模块

Fig. 5　Relabeling module

1.4　网络损失函数方案

现实中的图像分类数据集常常存在噪声标签，这些噪声标签可能呈现多模态特性。为了学习噪声

标签与干净标签之间的映射关系，研究者们引入了先验信息[7]，但端到端的优化方式可能导致损失函数降为零。为了解决这一问题，一些方法采用了交替最小化权重的策略[8-9]，然而这会增加时间成本。近年来，加权学习方法受到了关注，其中对数加权[10]结合了每个样本的对数和注意力值，更好地凸显了不确定性因素对损失函数的贡献。交叉熵损失函数作为图像分类的传统损失函数，在最大化不同类别标签之间的距离的同时最小化相同类别标签之间的距离。基于这一思想，本文采用了权重交叉熵损失函数（weight CE-loss，WCE-Loss），其公式为 $\mathrm{WCE\text{-}Loss} = -\frac{1}{N}\sum_{i=1}^{N} a_i y_i \ln(p_i)$，以更好地凸显了不确定性因素对损失函数的贡献。其中：N 表示样本数量，y_i 表示真实标签，p_i 表示预测概率，a_i 表示每个样本的权重。同时通过向损失函数添加一个基于边界偏置的正则化项，即 RR-Loss。这一项确保了高权重组的平均注意力权重值高于预设阈值的低权重组。在整个训练过程中，同时优化分类损失和正则化损失，其中 RR-Loss 强化了注意力机制的实际效果。RR-Loss 的公式表示为 $\mathrm{RR\text{-}Loss} = \max(0, \mathrm{margin} - \bar{a}_{\mathrm{H}} + \bar{a}_{\mathrm{L}})$。其中：W 是权重均值；margin 是边界阈值。

2 实验结果与分析

2.1 环境设置

本文采用了一种端到端的训练方法，并利用 1 块 NVIDIA Tesla V100 32 GB GPU 进行高效训练。为了确保模型能够充分学习到数据的特征，将每批大小设置为 512，并根据 β 超参数将训练图像分成了高权重组和低权重组，增强模型对不同数据的适应性。在损失函数方面，采用了 RR-Loss 和 WCE-Loss 相结合的优化策略，促进模型在训练过程中更好地收敛。为了进一步提升训练效果，将学习率初始化为 0.001，并通过权重衰减逐渐减小学习率，确保模型在训练后期能够更好地适应数据。同时，从第 10 轮训练开始启用了标签再分配模块，提高模型对标签的利用率。在数据增强方面，采用了多种策略对图像数据进行增强，包括随机水平翻转、随机裁剪和颜色抖动，以增加训练样本的多样性，提升模型的泛化能力。此外，还通过调整图像的亮度、对比度和饱和度等参数来模拟不同的拍摄条件，增强模型对光照变化的鲁棒性。

为了进一步提升效果，本文采用了多种训练策略。具体来说，使用了余弦退火学习率衰减，并在前 10 个 epoch 保持不变，之后采用余弦曲线衰减学习率，最后 5 个 epoch 将学习率线性衰减至 1×10^{-6}，以更加平滑地调节学习率，避免训练过程中的剧烈学习率变化。同时，对全连接层权重矩阵添加了 L2 正则项，并将正则化系数设为 1×10^{-4}，以惩罚模型权重，抑制过拟合。考虑到 GPU 显存和算力的限制，使用了 FP16 半精度数据格式进行模型训练，提高训练效率，使模型更容易收敛，提高运算速度。此外，还采用了标签平滑技术，将 0.1 的均匀分布噪声混合到原始 one-hot 标签中，以增强模型对小扰动的鲁棒性。为了加速模型的收敛速度，在前 5 个 epoch 线性增加学习率至最高峰 0.1，并使用了学习率预热技术，避免训练初期的不稳定问题。最后，使用了 Focal Loss 来代替交叉熵损失，并将平衡因子设定为 0.25，降低易分样本的权重，使模型更加关注难分样本。整体的参数设置如表 1 所示。

表 1 模型训练实验参数表

Tab.1 Model training experiment parameter table

类别	项目	描述
硬件环境	GPU	1 块 NVIDIA Tesla V100 32GB
训练数据	批大小	512
训练数据	训练数据划分	70% 确定样本;30% 不确定样本
模型结构	损失函数	RR-Loss + WCE-Loss
优化超参数	初始学习率	0.001
优化超参数	权重衰减	最后 5 个 epoch 线性衰减至 1×10^{-6}
优化超参数	余弦退火	前 10 个 epoch 固定 0.1,之后余弦衰减
训练策略	学习率预热	前 5 个 epoch 线性增长学习率至 0.1
数据增强	数据增强	随机水平翻转,随机裁剪,颜色抖动等
数据增强	图像增强	变更图像亮度、对比度、饱和度等

2.2　数据集

RAF-DB 面部表情数据集是一个广泛应用于情感计算和人脸识别领域的大型数据库。它包含了 29672 张真实世界的面部表情图像，这些图像都是在不同光照条件下、不同角度和不同面部表情强度下拍摄的，从而保证了数据的多样性和真实性。对于每一张图像，RAF-DB 都提供了详细的标注信息，包括表情强度、表情类型和面部关键点坐标等，有助于更深入地了解人类面部表情的表达方式和识别方法。特别是在单标签子集中，RAF-DB 包含了 7 类基本情绪：中性表情、震惊、恐惧、愤怒、高兴、失望和难过。这些情绪是人类日常生活中最为常见和最基本的情感表达方式，也是情感计算领域研究的重要对象。通过对这些基本情绪的深入分析和研究，有助于更好地理解和识别人类的情感状态，进而应用于人机交互、智能机器人、心理健康等领域。RAF-DB 数据集样本如图 6 所示。

MMAFEDB 数据集包含用于训练、验证和测试的文件目录。每个目录包含对应于七种面部表情类别的七个子目录。超过 6 万张各类表情的图像，在基于 Opencv 对人脸进行对齐后，截为 48 ×48 像素小尺寸的图像，然而并没有对面部图像进行人脸截取，所以会保留更多的无关信息和存在更多的不确定性。因此对模型泛化能力的测试更具挑战性，也更能凸显模型对非实验室环境下的面部表情识别性能的泛化。

图 6　RAF-DB 数据集

Fig.6　RAF-DB dataset

SMIC-E 数据集是 SMIC 数据集的升级版，它专门用于微表情检测研究。该数据集提供了更长的视频片段，其中包含非微帧的内容，为研究者提供了更丰富的上下文和情感信息。为了满足不同需求和环境条件，SMIC-E 数据集包含了三个数据子集，分别是使用高帧率相机捕捉的 HS 子集、使用普通相机拍摄的 VIS 子集，以及使用近红外线相机消除光照影响的 NIR 子集。每一个样本都被详细标注，分为正向表情、负面表情和激动表情三种基本情感类型。这些标签经过专业心理学家的标注，确保数据的准确性和可靠性。综上所述，SMIC-E 数据集为微表情研究领域提供了全面、多样化的数据来源，有助于研究者更深入地探究人类的微妙情感表达。

CASMEII 是 CASME 数据集的升级版，相较于原始版本，它在技术和细节上有了显著的提升。首先，CASMEII 采用了更高的帧率，达到了 200 fs，这使得微小的表情变化能够更加精细地被捕捉和记录。其次，CASMEII 提供了更大的人脸尺寸，分辨率为 280 ×340 像素，这样的尺寸确保了人脸表情的清晰度和细节可见度，从而提高了数据的质量和研究的准确性。CASMEII 的样本标签涵盖了 5 种情

感类型，包括开心、烦躁、惊讶、厌恶和其他。这些标签不仅基于面部表情，还结合了情境和上下文信息，确保情感标注的准确性和全面性。这些细致的标签为研究者提供了更广泛的研究领域和更深入的分析角度，有助于更精确地解读和理解人类的微表情及其背后的情感。

EmotioNet 数据集是一个用于面部表情识别的大规模图片数据集，常被用作预训练数据集，它包含了超过 95 万张来自互联网的人脸图片，涵盖了丰富的表情表达，包括基本表情和复合表情，以及表情单元的标注。EmotioNet 数据集的目的是为了帮助开发和评估面部表情识别算法，尤其是在真实世界中的复杂和多样化的场景下。

2.3 超参数选值

在超参数调优的过程中，对四个主要的超参数进行了详尽的探索，它们分别是 δ_1（margin）、δ_2（m_2）、β 和损失函数权重系数（weight），如图 7 所示。δ_1 作为正则化排序模块中的损失函数阈值，决定了高低权重组之间的距离何时开始产生损失。δ_2 的取值在标签再分配模块中决定了敏感度。通过计算再分配模块计算所得的最大预测概率与原模型输出标签对应的预测概率的距离，并与 δ_2 进行比较，以此决定最终的标签分配。通过大量的实验和迭代，本文得以在 RAF-DB 数据集上找到了它们的最佳取值。具体来说，当 δ_1 取值为 0.24，δ_2 取值为 0.27，β 取值为 0.4，weight 取值为 0.5 时，本文的模型表现出了最佳的精度。

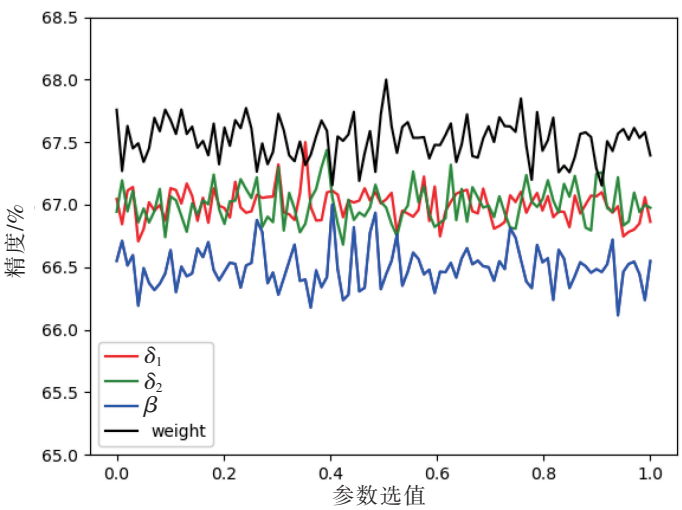

图 7 RAD-DB 数据集在不同超参数选值下的精度

Fig.7 Accuracy of RAD-DB under different hyperparameter settings

2.4 LUSLPA 在视觉模型上的表现

表 2 对比分析了四个表情识别数据集（RAF-DB、MMAFEDB、SMIC-E、CASME II）上，SOTA 模型在引入 LUSLPA 前和后的性能差异。数据清晰地指出，引入 LUSLPA 显著提高了各模型在表情识别任务上的表现。例如，在 RAF-DB 数据集中，集成 LUSLPA 后，ResNet、MobileNetv3、ShuffleNetv2 及 EfficientNetv2 等模型的准确率分别提高了 8.91%、6.09%、5.15%、4.89%。此类性能提升在其他 3 个数据集中也有所体现，证明了 LUSLPA 模块的高度泛化能力，能够有效地适配不同规模和类型的神经网络模型，从而增强其学习表情特征的能力。

在具体的参数量和计算量上，LUSLPA 所带来的增加非常少。以 EfficientNetv2 为例，算法的整合仅增加了 0.6 M 的参数和 20 M 的计算量，便实现了 4.89% 的准确率增幅，这一数据有效地展示了 LUSLPA 的高效性。对于像 ShuffleNetv2 这样的轻量级模型，尽管参数量和计算量分别只增加了 6% 和 10%，在 RAF-DB 数据集上的性能仍获得了 5.15% 的提升，显示出 LUSLPA 特别适合于计算资源受限的环境。

此外，LUSLPA 对性能提升的贡献程度与模型的规模及类型密切相关。对于参数和计算资源消耗较大的模型，如 RelCol 和 CageViT，尽管它们已具备较强的表情特征表示能力，LUSLPA 的集成依旧为其带来约 1% 的性能提升。这一现象进一步验证了 LUSLPA 的注意力融合策略具有普适性和有效性。

表 2　添加 LUSLPA 后各基线视觉模型在表情识别数据集指标上的提升

Tab. 2　Improvement of baseline visual models on facial expression recognition dataset after adding LUSLPA

模型	参数量/M	计算量/M	RAF-DB 精度/%	MMAFEDB 精度/%	SMIC-E 精度/%	CASMEII 精度/%
ResNet[11]	5.7	216	53.30	76.32	45.71	77.96
+ LUSLPA	6.3	236	8.91 ↑	2.88 ↑	5.60 ↑	4.20 ↑
MobileNetv3[12]	10.7	278	65.30	72.36	50.70	78.71
+ LUSLPA	11.3	298	6.09 ↑	2.73 ↑	4.57 ↑	4.46 ↑
ShuffleNetv2[13]	6.8	189	68.03	69.33	47.54	76.63
+ LUSLPA	7.4	209	5.15 ↑	3.34 ↑	3.20 ↑	1.13 ↑
EfficientNetv2[14]	8.3	1010	62.10	84.46	58.31	82.89
+ LUSLPA	8.9	1030	4.89 ↑	2.28 ↑	2.83 ↑	1.96 ↑
LiTv2[15]	10.4	152	58.00	77.15	48.49	63.96
+ LUSLPA	11.0	172	6.63 ↑	2.80 ↑	5.09 ↑	12.02 ↑
PVTv2[16]	8.8	158	51.47	78.26	53.95	81.10
+ LUSLPA	9.4	178	5.72 ↑	4.14 ↑	4.38 ↑	1.09 ↑
SepViT[17]	6.7	769	64.30	85.89	58.05	76.48
+ LUSLPA	7.3	789	5.52 ↑	2.01 ↑	1.43 ↑	2.24 ↑
Swin[18]	7.1	398	54.90	80.21	50.15	80.11
+ LUSLPA	7.8	418	6.57 ↑	4.27 ↑	6.46 ↑	5.76 ↑
CageViT[19]	17.6	285	65.21	87.34	61.65	79.57
+ LUSLPA	18.2	305	4.93 ↑	0.10 ↑	0.72 ↑	2.39 ↑
RelCol[20]	60.0	1104	69.19	84.29	58.58	83.96
+ LUSLPA	60.6	1124	1.29 ↑	1.12 ↑	1.16 ↑	1.19 ↑
EMO[21]	15.3	603	63.48	86.48	60.28	82.18
+ LUSLPA	15.9	623	3.17 ↑	0.35 ↑	1.62 ↑	1.44 ↑
SwiftFormer[22]	12.1	447	67.29	85.46	60.37	84.72
+ LUSLPA	12.7	467	1.71 ↑	1.16 ↑	1.05 ↑	1.35 ↑

注：↑表示提升率。

2.5　不同降噪方法在含有噪音的各个数据集上的评估

本节针对在不同噪声背景下，EfficientNetv2 模型结合不同降噪方案的性能进行了深入的比较和分析。表 3 展示了在无预训练情况下和利用 EmotioNet 预训练的情况下，LUSLPA 方案与 CurriculumNet 和 MetaCleaner 在不同噪声水平上的表现对比。从数据可见，LUSLPA 模块不仅在有无预训练的条件下均显示出了明显优势，其在抑制噪声方面的性能尤为突出。具体来说，当数据集噪声水平为 10% 时，LUSLPA 在 RAF-DB 数据集上实现了 3.70% 的准确率提升，相比之下，CurriculumNet 和 MetaCleaner 的提升率分别为 2.34% 和 2.72%。这一对比充分显示了 LUSLPA 在处理低噪声环境下的高效性能。

随着噪声水平的增加，LUSLPA 抑制噪声的能力愈加明显。例如，在噪声水平达到 30% 时，LUSLPA 在四个不同的表情识别数据集上的平均性能提升分别为 5.05%、2.03%、3.29% 和 2.44%。这些数

据不仅证明了 LUSLPA 的强大抗噪声能力，同时也突显了其在高噪声环境下依然能够维持较高的识别准确率。此外，LUSLPA 在经过 EmotioNet 预训练后，其性能进一步得到了增强。在 10% 噪声水平下，与未采用噪声抑制方案相比，预训练后的 LUSLPA 在各数据集上的性能增益介于 2.00% 至 5.01% 之间。这一结果表明，预训练结合 LUSLPA 能有效提升模型对关键表情噪声的识别准确性和可靠性。

表 3 EfficientNetv2 在含有噪音的四个数据集上使用不同降噪方案的评估

Tab. 3 Evaluation of efficientNetv2 on various datasets with different denoising strategies for noisy data

预训练	抑制方案	噪声	RAF-DB 精度/%	MMAFEDB 精度/%	SMIC-E 精度/%	CASMEII 精度/%
无预训练	无	10	60.41	81.59	52.73	78.84
	CurriculumNet[23]	10	2.34 ↑	1.51 ↑	3.42 ↑	2.46 ↑
	MetaCleaner[24]	10	2.72 ↑	1.84 ↑	5.35 ↑	2.35 ↑
	LUSLPA	10	3.70 ↑	1.55 ↑	4.98 ↑	3.42 ↑
	无	20	58.50	77.57	51.62	74.95
	CurriculumNet	20	4.54 ↑	1.16 ↑	3.72 ↑	2.38 ↑
	MetaCleaner	20	3.31 ↑	2.62 ↑	3.52 ↑	2.74 ↑
	LUSLPA	20	5.43 ↑	2.67 ↑	4.89 ↑	3.30 ↑
	无	30	54.71	72.78	49.53	69.56
	CurriculumNet	30	4.49 ↑	1.18 ↑	2.58 ↑	1.51 ↑
	MetaCleaner	30	4.95 ↑	1.60 ↑	1.61 ↑	1.82 ↑
	LUSLPA	30	5.05 ↑	2.03 ↑	3.29 ↑	2.44 ↑
EmotioNet 预训练	无	10	64.52	83.93	54.79	81.44
	CurriculumNet	10	1.25 ↑	1.16 ↑	3.18 ↑	3.12 ↑
	MetaCleaner	10	1.69 ↑	1.32 ↑	3.54 ↑	3.33 ↑
	LUSLPA	10	2.00 ↑	3.23 ↑	5.01 ↑	3.31 ↑
	无	20	62.72	80.65	53.48	78.86
	CurriculumNet	20	2.88 ↑	2.51 ↑	1.33 ↑	2.57 ↑
	MetaCleaner	20	3.35 ↑	1.50 ↑	1.46 ↑	2.48 ↑
	LUSLPA	20	4.38 ↑	2.45 ↑	1.93 ↑	2.63 ↑
	无	30	59.68	76.60	52.79	74.55
	CurriculumNet	30	2.18 ↑	2.72 ↑	2.05 ↑	1.75 ↑
	MetaCleaner	30	2.79 ↑	2.26 ↑	1.36 ↑	1.64 ↑
	LUSLPA	30	4.79 ↑	3.23 ↑	1.65 ↑	2.64 ↑

注：↑表示提升率。

2.6 对比 SOTA 模型

表 4 详细展示了 LUSLPA 模型在四个不同的表情识别数据集（RAF-DB、MMAFEDB、SMIC-E、CASME II）上的性能，并与其他几种先进的模型进行了比较。其中，IPA2LT[25] 引入了潜在的真实性思想，可在不同 FER 数据集之间使用不一致的注释进行训练。gaCNN[26] 利用基于 token 的注意力网络和全局网络。RAN[27] 利用人脸区域和原始人脸并具有级联注意力网络。由于裁剪了 token 和区域，因此 gaCNN 和 RAN 非常耗时。而 LUSLPA 不会增加任何推理成本。从表 4 中数据可以清晰观察到，LUSLPA 模型在所有数据集上均展现出较其他模型更高的精度，具体来说，LUSLPA-ResNet18 在 RAF-

DB 上的精度达到了 63.21%，比目前已知最高精度的 PLD 模型高出 1.07%；在 MMAFEDB 上达到了 61.89%；在 SMIC-E 数据集上，精度为 62.77%，较第二高的 RAN-VGG16 模型的 62.13% 还高出 0.64%；在 CASME II 数据集上，则以 60.09% 的精度领先，高 Weighted Loss 模型 58.71%。

表 4　LUSLPA 与其他 SOTA 模型的对比

Tab. 4　Comparison of LUSLPA with other SOTA models

方法	RAF-DB 精度/%	MMAFEDB 精度/%	SMIC-E 精度/%	CASMEII 精度/%
DLP-CNN[28]	54.22	52.80	53.90	51.75
IPA2LT	60.45	58.61	57.90	56.40
gaCNN	59.80	57.75	58.40	55.83
RAN	61.70	59.10	60.00	58.25
RAN-VGG16	61.72	59.07	62.13	57.08
Upsampler[29]	60.80	59.40	60.33	58.57
Weighted Loss[29]	61.32	60.08	60.75	58.71
PLD[30]	62.14	60.89	61.27	59.40
ResNet + VGG[11]	56.83	59.96	55.62	58.91
SeNet50[31]	57.37	60.18	57.49	58.70
LUSLPA-ResNet18	63.21	61.89	62.77	60.09
LUSLPA-IR50	64.32	62.07	62.84	61.00

2.7　消融实验

表 5 展示了在 RAF-DB 数据集上，使用不同模块组合后的 EfficientNetv2 在表情识别任务上的性能。可明显看出，引入本文提出的自注意力加权模块、正则化排序模块和标签再分配模块，模型的识别准确率得到明显提升。具体来说，仅使用 WCE-Loss 作为损失函数时，普通卷积模块作为特征提取器的模型精度为 61.24%；当将其中部分卷积块替换为 Self-Attention 模块或动态卷积模块后，精度略有提升，分别达到 61.79% 和 62.28%。这表明注意力机制对学习表情特征有一定帮助。当加入标签再分配模块后，所有三种特征提取器的精度均有不同程度的提升，尤其是动态卷积模块，从 62.28% 提升至 63.87%。这证明了标签再分配模块通过修正噪声标签，增强了模型的泛化能力。在此基础上，使用正则化排序模块后，精度将进一步大幅提升，尤其是动态卷积模块提升到 66.70%，说明正则化排序模块通过对样本集进行权重区分和约束有助于模型学习。

同时使用 WCE-Loss 和 RR-Loss 进行模型优化时，动态卷积模块的精度可进一步达到 66.99%，总体提升 5.75%。双损失函数使模型在特征表示和判别上都得到加强。可见，本文的模块的引入对精度提升具增量效果，模块组合的协同作用达到最优。这为表情识别提供了有效方向。

表 5　消融实验

Tab. 5　Ablation Study

自注意力加权模块	正则化排序模块	标签再分配模块	损失函数	精度/%
普通卷积模块			WCE-Loss	61.24
Self-Attention 模块			WCE-Loss	61.79
动态卷积模块			WCE-Loss	62.28
普通卷积模块		+	WCE-Loss	62.54
Self-Attention 模块		+	WCE-Loss	63.35
动态卷积模块		+	WCE-Loss	63.87

续表

自注意力加权模块	正则化排序模块	标签再分配模块	损失函数	精度/%
普通卷积模块	+	+	WCE-Loss	64.91
Self-Attention 模块	+	+	WCE-Loss	65.06
动态卷积模块	+	+	WCE-Loss	66.70
普通卷积模块			WCE-Loss + RR-Loss	61.59
Self-Attention 模块			WCE-Loss + RR-Loss	61.63
动态卷积模块			WCE-Loss + RR-Loss	62.75
普通卷积模块		+	WCE-Loss + RR-Loss	63.14
Self-Attention 模块		+	WCE-Loss + RR-Loss	63.38
动态卷积模块		+	WCE-Loss + RR-Loss	65.27
普通卷积模块	+	+	WCE-Loss + RR-Loss	65.94
Self-Attention 模块	+	+	WCE-Loss + RR-Loss	66.38
动态卷积模块	+	+	WCE-Loss + RR-Loss	66.99

注：+ 表示加入。

3 结论

本文开拓性地提出自监督标签不确定性预测分配算法，有效应对面部表情识别领域中干扰模型性能和鲁棒性的关键挑战。该算法创新三模块架构，针对性地缓解了标签不确定性的影响。自注意力加权模块灵活运用动态卷积，实现像素级精细关注机制，精准特征提取，同时降低计算复杂度；正则化排序模块优化不确定样本处理，通过样本权重重排增强不确定样本的鉴别；标签再分配模块校正低权重样本标注，实现标签的自学习算法。

通过广泛公开数据集的实证评估证明，该算法在消除标签不确定性影响方面表现卓越，凸显了推进该领域发展的巨大潜能，为构建鲁棒实用模型铺就坚实途径。后续研究可将算法与先进深度学习架构融合，扩大适用范畴；在规模庞大、多样化数据集上考察，为算法完善及泛化奠定基础。

[参 考 文 献]

[1] NEWMARK C, CHARLES D. The expression of the emotions in man and animals[M]//SENGE K, SCHÜTZEICHEL R, ZINK V. Schlüsselwerke der Emotionssoziologie. Wiesbaden: Springer Fachmedien Wiesbaden, 2022: 111-115.

[2] LIN J S C, HSIEH P. The role of technology readiness in customers' perception and adoption of self-service technologies[J]. International Journal of Service Industry Management, 2006, 17(5): 497-517.

[3] VALSTAR M, PANTIC M. Induced disgust, happiness and surprise: an addition to the MMI facial expression database[C]// Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect. Paris: ELRA, 2010: 65-70.

[4] ZHAO G, HUANG X, TAINI M, et al. Facial expression recognition from near-infrared videos[J]. Image and Vision Computing, 2011, 29(9): 607-619.

[5] CHEN Y, DAI X, LIU M, et al. Dynamic convolution: attention over convolution kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 11030-1039.

[6] YI X, YANG J, HONG L, et al. Sampling-bias-corrected neural modeling for large corpus item recommendations[C]//Proceedings of the 13th ACM Conference on Recommender Systems. New York: ACM, 2019: 269-277.

[7] HU W, HUANG Y, ZHANG F, et al. Noise-tolerant paradigm for training face recognition CNNs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 11887-11896.

[8] MINAEE S, MINAEI M, ABDOLRASHIDI A. Deep-emotion: facial expression recognition using attentional convolutional network[J]. Sensors, 2021, 21(9): 3046.

[9] MA F, MENG D, XIE Q, et al. Self-paced co-training[C]//Proceedings of the 34th International Conference on Machine Learning. New York: PMLR, 2017: 2275-2284.

[10] ALP GULER R, TRIGEORGIS G, ANTONAKOS E, et al. Densereg: fully convolutional dense shape regression in-the-wild [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 6799-6808.
[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 770-778.
[12] HOWARD A, SANDLER M, CHU G, et al. Searching for MobileNetV3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 1314-1324.
[13] MA N, ZHANG X, ZHENG H T, et al. Shufflenet v2: practical guidelines for efficient CNN architecture design[C]//Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 116-131.
[14] TAN M, LE Q. Efficientnetv2: smaller models and faster training[C]//Proceedings of the International Conference on Machine Learning. New York: PMLR, 2021: 10096-10106.
[15] PAN Z, CAI J, ZHUANG B. Fast vision transformers with hilo attention[J]. Advances in Neural Information Processing Systems, 2022, 35: 14541-14554.
[16] WANG W, XIE E, LI X, et al. Pvt v2: improved baselines with pyramid vision transformer[J]. Computational Visual Media, 2022, 8(3): 415-424.
[17] LI W, WANG X, XIA X, et al. Sepvit: separable vision transformer[EB/OL]. (2023-06-15)[2023-12-15]. https://doi.org/10.48550/arXiv.2203.15380.
[18] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 10012-10022.
[19] ZHENG H, WANG J, ZHEN X, et al. CageViT: convolutional activation guided efficient vision transformer[EB/OL]. (2023-05-17)[2023-12-15]. https://doi.org/10.48550/arXiv.2305.09924.
[20] CAI Y, ZHOU Y, HAN Q, et al. Reversible column networks[EB/OL]. (2023-02-01)[2023-12-15]. https://doi.org/10.48550/arXiv.2212.11696.
[21] ZHANG J, LI X, LI J, et al. Rethinking mobile block for efficient attention-based models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 1389-1400.
[22] SHAKER A, MAAZ M, RASHEED H, et al. Swiftformer: efficient additive attention for transformer-based real-time mobile vision applications[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 17425-17436.
[23] GUO S, HUANG W, ZHANG H, et al. Curriculumnet: weakly supervised learning from large-scale web images[C]//Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 135-150.
[24] ZHANG W, WANG Y, QIAO Y. Metacleaner: learning to hallucinate clean representations for noisy-labeled visual recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. New York: IEEE, 2019: 7373-7382.
[25] ZENG J, SHAN S, CHEN X. Facial expression recognition with inconsistently annotated datasets[C]//Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 222-237.
[26] LI Y, ZENG J, SHAN S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450.
[27] WANG K, PENG X, YANG J, et al. Region attention networks for pose and occlusion robust facial expression recognition [J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.
[28] LI S, DENG W, DU J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 2852-2861.
[29] 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(1): 18-31.
[30] BARSOUM E, ZHANG C, FERRER C C, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction. New York: ACM, 2016: 279-283.
[31] ALBANIE S, NAGRANI A, VEDALDI A, et al. Emotion recognition in speech using cross-modal transfer in the wild[C]//Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM, 2018: 292-301.

(责任编辑　彭海滨　英文审校　黄振坤)