DOI:10. 13880/j. cnki. 65-1174/n. 2024. 23. 007 文章编号:1007-7383(2024)01-0126-07

融合 PVT 多级特征的口罩人脸识别研究

冉瑞生[1],高天宇[1],房斌[2]
(1 重庆师范大学计算机与信息科学学院,重庆 401331;2 重庆大学计算机学院,重庆 400044)

摘要:呼吸系统疾病的流行使口罩扮演着重要角色,这给人脸识别算法带来了新的挑战。受到多尺度特征融合模型的启发,提出一种基于金字塔视觉 Transformer (Pyramid Vision Transformer, PVT)的提取口罩人脸特征的模型。该模型引入自注意力机制来提取丰富的人脸信息,通过融合 PVT 多个层级的特征向量,来实现对口罩人脸的多尺度关注,相较于传统特征融合模型,具有更高的识别精度和更少的参数量。此外,模型采用 Sub-center ArcFace 损失函数来提升鲁棒性。模型在大规模模拟口罩人脸数据集上进行训练,并分别在普通人脸、模拟口罩人脸和真实口罩人脸数据集上进行了测试和评估。实验结果表明,所提出的方法与其他主流方法相比,具有较高的识别精度,是一种有效的口罩人脸识别方法。

关键词:口罩人脸识别;Transformer;自注意力机制;特征融合
中图分类号:TP391. 41 **文献标志码**:A

Research on masked face recognition by fusing multi-level features of PVT

RAN Ruisheng[1], GAO Tianyu[1], FANG Bin[2]
(1 College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China;
2 College of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract: The prevalence of respiratory diseases has made masks play an important role, which has brought new challenges to face recognition algorithms. Inspired by the multi-scale feature fusion model, a Pyramid Vision Transformer (PVT) based face mask feature extraction model is proposed. The model introduces self-attention mechanism to extract rich face information, and realizes multi-scale attention to mask faces by fusing multi-level feature vectors of PVT. Compared with traditional feature fusion model, the model has higher recognition accuracy and fewer parameters. In addition, the model adopts Sub-center ArcFace loss function to improve robustness. The model was trained on a large scale simulated mask face dataset, and tested and evaluated on ordinary face, simulated mask face and real mask face dataset respectively. The experimental results show that the proposed method has higher recognition accuracy than other mainstream methods, and is an effective mask face recognition method.

Key words: masked face recognition; Transformer; self-attention; feature fusion

近年来,随着人工智能技术的不断发展,人脸识别技术已经被广泛应用于各个领域。然而,在当前全球呼吸系统疾病流行的背景下,佩戴口罩已成为一种必要的防护措施[1]。口罩对人脸的遮挡给人脸识别技术带来了新的挑战,成为降低准确率的主要原因之一。因此探索一种提取人脸鲁棒性特征的方法具有重要意义。当前,已经有部分口罩人脸识别算法被提出,例如,Mandal 等[2]利用 Resnet-50 模型,对未佩戴口罩人脸数据进行训练后再迁移到口罩人脸数据,旨在通过对未佩戴口罩的人脸数据进行训练,实现对口罩人脸的识别。姜绍忠等[3]提出一种 CNN 与 Transformer 相结合的混合模型,在人工合成的口罩人脸数据集上进行训练,所训练的模型能同时处理戴口罩和不戴口罩的人脸识别任务,但该方法缺乏对真实口罩人脸的验证。Li 等[4]提出一种基于裁剪和注意力机制的口罩人脸识别方法,该方法通过对人脸图像进行裁剪,以此来移除受损区域或降低遮罩区域的权重,并结合注意力机制来关注眼睛周围区域。这种方法能够更加有效地捕捉人脸的局部特征信息,从而提高模型的识别准确率。

收稿日期:2023-10-18
基金项目:重庆市教育委员会科学技术研究项目(KJZD-K202100505)
作者简介:冉瑞生(1976—),男,教授,从事机器学习方向的研究,e-mail: rshran@ cqnu. edu. cn。

然而,该方法会降低无口罩人脸识别的准确率。Qian 等[5]提出了一种方法,将 ArcFace 损失函数和 pairwise loss 结合起来,以增强遮挡人脸识别任务的性能。该方法旨在提高同一类别内样本的相似度,同时增加不同类别之间的差异性,从而提高遮挡人脸识别的准确性。

这些方法虽然能实现口罩人脸识别,但还是存在一些问题。首先,现有的大部分方式通过单一尺度特征进行预测,这样可能会忽略一些其他尺度的特征,例如,对于人脸而言,同时考虑眼睛大小和整个人脸轮廓的多尺度特征对于全面捕捉人脸特征至关重要。其次,当前的主流特征融合方法主要集中在特征图的整合上,这可能会增加计算负担。

针对以上问题,本文提出一种融合 PVT 各尺度特征的口罩人脸表征方法,该方法可同时用于佩戴口罩和不佩戴口罩的人脸识别场景。主干网络使用基于 MSA (Multi-head Self-Attention)改进的 PVT (Pyramid Vision Transformer)提取人脸的多尺度特征。在每个尺度阶段都使用 1 个 *cls* (*class token*)向量来存储该尺度的人脸特征,并通过融合各尺度的 *cls* 以使得提取的特征更加丰富。最后,使用 Sub-center ArcFace 损失函数来进一步提高模型的鲁棒性。该方法使用多个数据集进行验证,涵盖了多种人脸场景。实验结果表明,本文方法能有效提高口罩人脸识别的准确率,同时特征融合的计算量也相对较低。

1　资料与方法

本文提出了融合 PVT 多级特征的口罩人脸识别模型,命名为 PVTFace。

设输入图像为三通道(RGB)彩色图像,图像尺寸为 112×112。PVTFace 模型首先将图像分割为 196 个不重叠的图像块,每个图像块会被转换为向量形式,得到 Patch Embedding,然后拼接 *cls* 向量并添加位置信息,*cls* 用以存储图像特征,方便后续阶段的计算。随后将 Patch Embedding 输入到多个堆叠的 Transformer Encoder 中进行计算得到相应的特征图,Transformer Encoder 中的注意力机制使用 MSA[6]。特征图再输入到下一个 Stage 进行采样。完成各 Stage 采样后,再将各 Stage 的 *cls* 进行融合。最终,将融合后的图像特征送入 Sub-center ArcFace 损失函数进行计算。

PVTFace 网络结构如图 1 所示,接下来将对所改进的模块进行详细阐述。

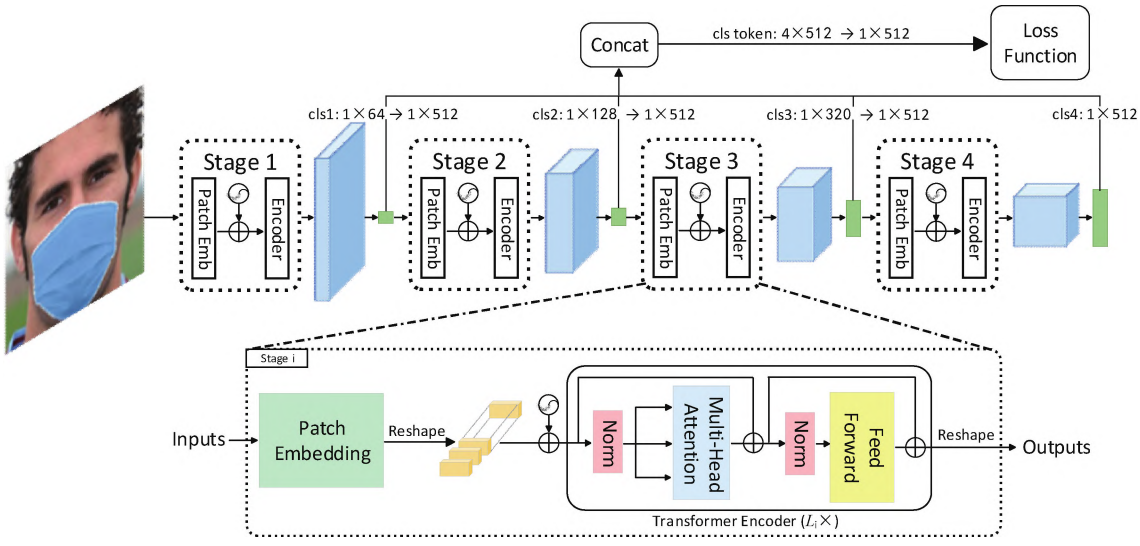

图 1　PVTFace 的网络结构

1.1　注意力机制

Spatial-Reduction Attention (SRA)[7]是 PVT 中提出的一种注意力机制,相较于 MSA,SRA 通过对键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$ 进行空间上的下采样,以达到降低计算复杂度的目的,SRA 与 MSA 的结构对比如图 2 所示。

然而,在口罩人脸识别的场景下,使用 SRA 对人脸图像进行下采样可能会导致忽略一些重要的特征。因为口罩遮盖了部分面部特征,如嘴巴、鼻子,所以降低空间分辨率可能会造成信息的丢失。在这种情况下,使用 SRA 可能会降低对于口罩人脸的识别准确性。

Self-Attention 可以在输入序列中建立长依赖关系,且能对输入序列中的所有位置进行关注,从而能够捕捉全局的语义信息。在人脸识别任务中,由于人脸图像中的各个部分之间存在较强的相关性。

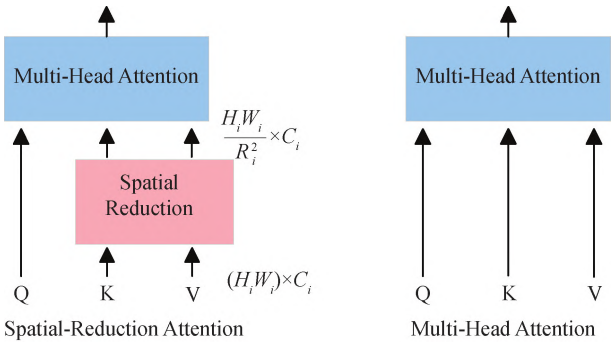

图 2 SRA 与 MSA 的结构对比

Self-Attention 可以有效地将这些关系建模,提高人脸识别的准确率。并且 Self-Attention 对于输入序列的变化(例如旋转、缩放、遮挡等)具有很强的适应性,因此可以提高模型的鲁棒性。其公式表述为:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}。 \quad (1)$$

其中, $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ 分别为查询、键、值,它们由神经网络训练得到。

传统的 Transformer 使用基于 Self-Attention 机制的 MSA。MSA 是 Self-Attention 的扩展形式,它通过使用多个注意力头来提供多个视角的关注能力。每个注意力头可以专注于不同的特征子空间或关系,从而捕捉到输入序列的不同方面和语义信息。通过融合多个头的结果,MSA 能够提供更全面和丰富的表示,进而增强模型对输入序列的建模能力。因此,本文使用 MSA 作为注意力模块,以便更好地捕捉序列的多样性特征和语义信息。

1.2 特征融合

以往的基于深度学习的人脸识别模型都过于注重深层次特征,即只使用网络的最后一层特征作为身份特征,这样可能会忽略浅层次的人脸特征[8]。在此基础上,本文提出一种基于 PVT 的人脸识别架构,通过融合各层次的特征来提取人脸的鲁棒性特征。

在每个 Stage 中,输入数据首先计算得到 Patch Embedding,随后通过 concat 方式拼接 1 个 *cls* 向量用于存储该 Stage 的特征信息,再输入到多个堆叠的 Transformer Encoder 中进行计算,4 个阶段的 *cls* 维度分别为 1×64,1×128,1×320,1×512。再将各 Stage 中的 *cls* 维度全部映射为 1×512,这样做的目的是为了保证各个 Stage 的特征信息可以得到充分的利用,并且各个特征具有相同的维度,便于后续的特征融合和计算,过程如图 1 所示。将各 Stage 的 *cls* 进行 concat 拼接得到维度为 4×512 的 *cls token*,具体的特征融合过程可以表示为:

$$\begin{aligned}
&\boldsymbol{cls1}: dim_{1\times 64} \rightarrow dim_{1\times 512},\\
&\boldsymbol{cls2}: dim_{1\times 128} \rightarrow dim_{1\times 512},\\
&\boldsymbol{cls3}: dim_{1\times 256} \rightarrow dim_{1\times 512},\\
&\boldsymbol{cls4}: dim_{1\times 512}。\\
&\boldsymbol{cls\quad token} = \boldsymbol{cls1} + \boldsymbol{cls2} + \boldsymbol{cls3} + \boldsymbol{cls4},\\
&\boldsymbol{cls\quad token}: dim_{4\times 512} \rightarrow dim_{1\times 512}。
\end{aligned} \quad (2)$$

式中,*dim* 表示 *cls* 的维度,→ 表示维度映射变化。随后将拼接得到的 *cls token* 的维度由 4×512 映射为 1×512,这样就使得 PVTFace 计算出的图像表征与原始 PVT 计算出的图像表征具有相同特征维度,却又包含了更加丰富的表征信息。

1.3 Sub-center ArcFace 损失函数

目前主流的深度人脸识别方法,如 CosFace[9]、ArcFace[10]在无约束的人脸识别中取得了显著的成功。然而这些方法通常只为每个类别设置一个中心,这种设计在受到噪声和变化的影响时可能会导致较差的鲁棒性。Sub-center ArcFace[11]为每个类别引入了 K 个子中心,训练样本只需要接近 K 个正向子中心中的任何一个。这样的设计可以更好地处理真实世界中的噪声和变化,提高模型的稳健性。

Sub-center ArcFace 具体实现方式是,为每个身份设置 1 个 K,并根据嵌入特征 $x_i \in R^{512\times 1}$ 和所有子中心 $W \in R^{N\times K\times 512}$ 进行归一化处理,通过矩阵相乘计算得到子类的相似得分 $S \in R^{N\times K}$,然后对子类相似度得分进行最大池化以得到类的相似度评分 $S' \in R^{N\times 1}$。Sub-center ArcFace 损失函数可以表述为:

$$\ell_{\text{ArcFace}_{\text{subcenter}}} = -\log\frac{e^{\text{scos}(\theta_{i,y_i}+m)}}{e^{\text{scos}(\theta_{i,y_i}+m)} + \sum_{j=1,j\neq y_i}^{N} e^{\text{scos}\theta_{i,j}}}。 \quad (3)$$

其中, $\theta_{i,j} = \arccos(\max_k(\boldsymbol{W}_{jk}^T x_i))$, $k \in \{1,\cdots,K\}$。

2 结果与分析

本文实验在 Linux 环境下进行,使用的 GPU 为单个 NVIDIA A100 PCIe,批量大小为 128,总 epoch 为 20,优化器为 AdamW,初始学习率为 3×10^{-4}。本节将介绍本文所使用的数据集及相关处理,并通过分析实验结果来验证本文所提方法的有效性。

2.1 数据集

MS-Celeb-1 M[12]是微软公司于 2016 年发布的

一个大规模人脸数据集,其中包含 400 万张照片和 79 057 个人物的标签信息。本文首先对 MS-Celeb-1 M 数据集进行清洗,再使用开源工具 MaskTheFace[13]来对该数据集中的人脸生成虚拟口罩,得到 MS-Celeb-1 M_masked,并以此作为训练集。

本文使用了多个测试集,分别为 LFW[14],LFW_masked,SLLFW[15],SLLFW _masked,CPLFW[16] 和 RMFD (Real-World Masked Face Dataset)[17]。其中 LFW 是由美国马萨诸塞州立大学阿默斯特分校计算机视觉实验室整理完成的数据库,包含 13 233 张照片和 5 749 个人物的标签信息;LFW_masked 是使用 MaskTheFace 对 LFW 人脸进行掩码处理后生成的口罩测试集。SLLFW 数据集是基于 LFW 实现,它构建了一组相似但非同一人的人脸对,该数据集旨在考察算法对于相似人脸的区分能力,提供更接近真实场景中的人脸验证情况。SLLFW_masked 是 MaskTheFace 对 SLLFW 人脸进行掩码处理得到的口罩数据集。CPLFW 数据集是在 LFW 基础上进行扩充,包含了多个姿态的人脸图像,如正脸、侧脸等,目的是提供更具挑战性的人脸验证场景。RMFD 是武汉大学国家多媒体软件技术研究中心开放的真实口罩人脸数据集,涵盖了 525 人的 5 000 张口罩人脸图像。部分数据集图像如图 3 所示,这些数据集包括正常人脸、模拟口罩人脸和真实口罩人脸数据,同时考虑了人脸姿态等多种场景,可更全面地评估模型的识别性能。

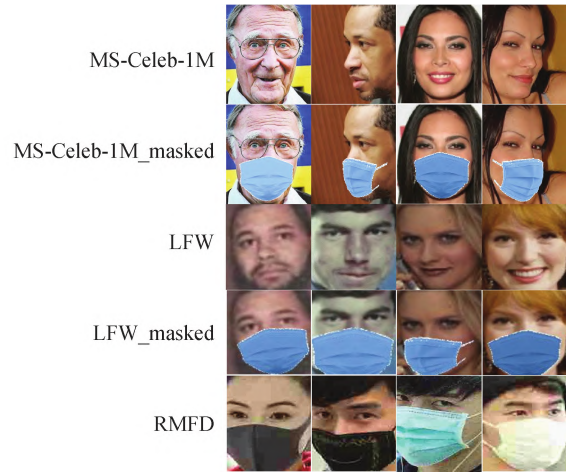

图 3　部分数据集示例

2.2　实验与分析

2.2.1　实验与分析

本文通过与 Resnet-50[18],Resnet-50f,GhostNet[19],MobileFaceNet[20],ViT[21] 以及 PVT 等模型进行对比,其中,ViT 和 PVT 是基于 Transformer 实现的模型。Resnet-50f 是基于 Resnet-50 实现特征融合模型,用于与基于 PVT 特征融合的 PVTFace 进行比较。以上与 PVTFace 对比的方法全部使用 CosFace 作为损失函数。评估结果如表 1 所示,PVTFace 在各测试集上的识别准确率均明显高于其他模型。这表明 PVTFace 在仅能提取少量特征的情况下进行训练就能兼顾戴口罩、不戴口罩、多姿态以及相似人脸区分等复杂情况。

表 1　不同方法在不同数据集上的比较和结果

Method	准确率/%					
	LFW	LFW_masked	SLLFW	SLLFW_masked	CPLFW	RMFD
Resnet-50	69.73	56.37	55.93	52.85	54.38	76.58
Resnet-50f	75.47	63.76	59.17	54.19	56.07	77.65
GhostNet	66.75	56.58	53.78	53.07	52.93	77.00
MobileFaceNet	62.32	57.56	55.65	56.48	53.55	76.53
ViT	58.28	54.45	53.57	53.91	52.25	77.36
PVT	70.25	61.72	54.45	53.47	51.60	76.44
PVTFace	**84.98**	**66.04**	**67.25**	**60.84**	**61.48**	**77.90**

值得注意的是,ViT 虽然在处理自然图像等领域表现优异,但其只关注深层次特征,在处理面部遮挡等复杂场景下存在局限性,因为其缺乏空间信息的连续性和不变性,难以充分捕捉面部的细节特征。而金字塔模型(PVTFace 和 Resnet-50f)可以使用不同的感受野来捕捉不同尺度的信息,包括全局尺度和局部尺度。在全局尺度上,模型可以识别人脸的大体特征,如整体轮廓和人脸区域的大小和形状。在局部尺度上,模型可以更加精细地识别人脸的细节特征,如眼睛、额头等部位。

为了更进一步评估模型的整体性能水平,本文以 LFW 数据集测试结果为基础绘制了上述各方法的 ROC 曲线进行对比分析。如图 4 所示,纵轴代表真阳性率(TPR),横轴代表假阳性率(FPR)。以 ROC 曲线下方的面积(AUC)来评价方法的优劣,由此可见,PVTFace 的识别效果远高于其他方法。

图 5 为本次实验 7 种模型在测试集 LFW 的准确率折线图,其中,基于特征融合实现的方法(如

PVTFace, Resnet-50f)的识别准确率明显高于其他方法。PVTFace 迭代四轮以后就能达到最佳效果。

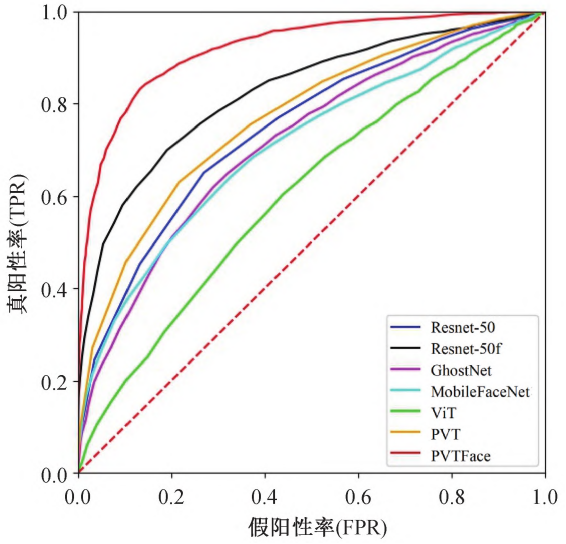

图 4 各模型的 ROC 曲线

表 2 展示了 PVT 使用不同特征融合方式识别准确率,各模型除了特征融合方式以外其他条件均一致。其中,*cls_add* 表示使用 add 相加的方式将各 Stage 的 *cls* 向量相加;AFF 表示使用基于注意力实现特征融合的 AFF (Attentional Feature Fusion)[22],将不同尺度的特征图进行融合;FPT 表示使用 FPT (Feature Pyramid Transformer)[23]所提出的特征增强方式对各尺度的特征图进行融合与增强;*cls_concat* 是本文所使用的特征融合方式,将各 Stage 的 *cls* 向量通过 concat 方式拼接。实验结果表明,*cls_add* 方式在 LFW_masked 数据集的识别率略高于本文方法,但在其他数据集上的验证并不如本文方法。AFF 方式在 RMFD 数据集的验证中取得了最佳结果,但由于其使用特征图融合的方式,参数量是本文方法的 2 倍多,提升效果却并不高。FPT 方式也会产生较多模型参数,且效果不佳。

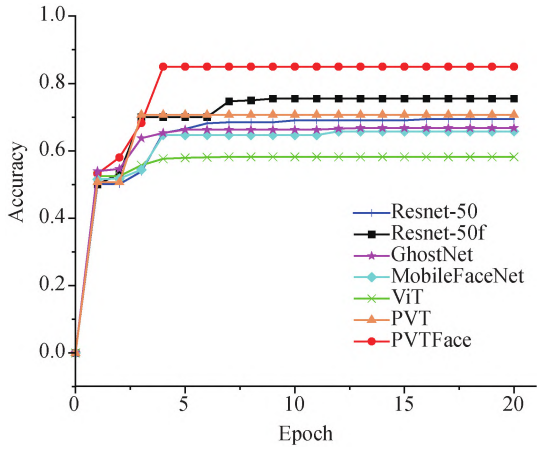

图 5 不同模型训练过程的准确率

表 2 不同特征融合方法在各数据集上的比较和结果

特征融合方式	参数/M	Accuracy/%					
		LFW	LFW_masked	SLLFW	SLLFW_masked	CPLFW	RMFD
cls_add	19.55	76.67	66.34	61.63	55.11	55.56	76.98
AFF	50.42	83.08	65.86	63.65	57.89	58.43	78.18
FPT	55.39	73.98	61.96	57.23	54.52	57.17	76.85
cls_concat	**20.60**	**84.98**	**66.04**	**67.25**	**60.84**	**61.48**	**77.90**

针对 Sub-center ArcFace 损失函数中子中心数量(K)对提取口罩人脸特征的影响,本文分别对 K 取值 1、3、5 在口罩人脸数据集上进行实验,结果如表 3 所示。观察发现,当 K 取值 3 时,在 3 个数据集中均取得了最优效果。这表明针对口罩人脸数据集,适当放宽数据的类内约束可以提高模型的鲁棒性。

表 3 Sub-center ArcFace 损失函数子中心数量(K)对口罩人脸识别的影响

K	Accuracy/%		
	LFW_masked	SLLFW_masked	RMFD
1	63.12	54.78	76.93
3	**66.04**	**60.84**	**77.90**
5	63.87	56.73	76.69

2.2.2 Grad-CAM 可视化

为了更加直观的分析实验结果,本文使用 Grad-CAM[24]生成类热力图,以此来可视化 Resnet-50, ViT, PVT, Resnet-50f 和 PVTFace 的注意力分布。如图 6,图中颜色越深代表此处模型权重越高,即模型更加关注该区域。PVTFace 各层关注点为面部轮廓、额头以及眼睛区域,将各层特征进行融合以后基本可以得到除口罩以外的所有面部区域。Resnet-50f 各层关注重点集中在额头区域,而忽略了眼睛部位和面部轮廓的信息。其他模型都只关注了局部面部信息,这也是这些方法准确率低的重要原因。

2.2.3 模型参数量与计算量分析

表 4 展示了 Resnet-50, ViT, PVT, Resnet-50f 和 PVTFace 的参数量(Params)和计算量(MACs)。其中,Resnet-50f 具有较多的参数量,这是因为在进行特征融合时它融合了整个特征图,而 PVTFace 仅融合了 *cls* 向量,从而大大减少了模型的参数量。

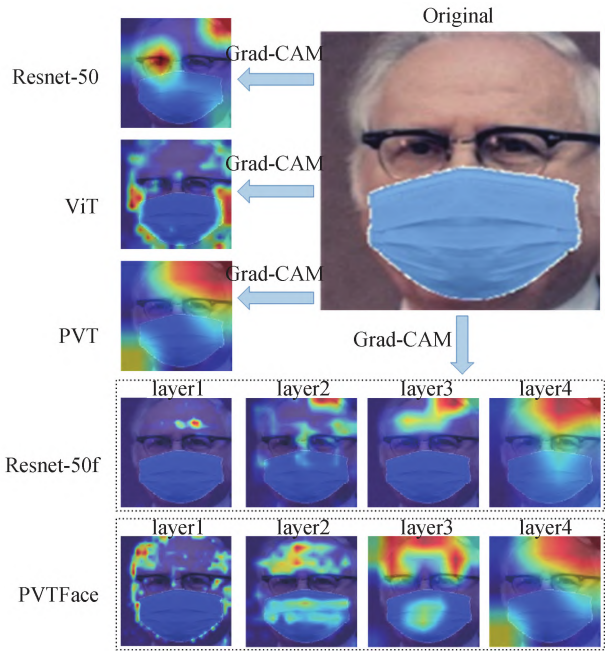

图 6　不同方法的 Grad-CAM 可视化的比较与结果

另外,ViT 是基于自注意力机制的柱状结构,因此导致其计算量较大。相比之下,PVTFace 相对于 PVT 仅增加了少量的模型参数和计算量,却取得了显著的识别效果,这表明所增加的参数量和计算量是值得的。此外,PVTFace 的参数量和计算量都小于 Resnet-50 模型,突显了所提出模型的优越性。

表 4　不同方法在不同数据集上的比较和结果

Method	Params/M	MACs/M
Resnet-50	27.70	1091.33
Resnet-50f	167.16	1231.06
ViT	63.12	12429.58
PVT	19.28	931.26
PVTFace	**20.60**	**941.51**

2.3　消融实验

本节通过在各测试集上进行消融实验来验证该方法的有效性,实验结果如表 5 所示。表 5 的第一列为模型名称,其中“+”表示在上一个模型基础上进行的改进。“+MSA”表示将基准模型(PVT)的注意力机制由 SRA 改为基于自注意力机制的 MSA;“+Sub-center”表示在上一个模型的基础上,将损失函数替换为 Sub-center ArcFace;“+Feature Fusion”表示在上一个模型的基础上,将各层特征进行融合。通过这些消融实验,证实了每个改进对模型性能的影响,并展示了提出方法的有效性。

根据实验结果,可以发现在使用 MSA 和 Sub-center ArcFace 损失函数后,模型的识别准确率有了显著提升。而在进行特征融合后,模型的识别率进一步提高。这表明所引入的 MSA、Sub-center ArcFace 损失函数以及特征融合操作对提升模型性能起到了积极的作用。

表 5　在不同数据集上的消融实验

Method	Accuracy/%				
	LFW	LFW_masked	SLLFW	SLLFW_masked	CPLFW
BaseLine	70.25	61.72	54.45	53.47	51.60
+MSA	78.73	63.49	64.37	56.45	56.63
+Sub-center	84.20	65.17	66.42	58.57	60.08
+ Feature Fusion	84.98	66.04	67.25	60.84	61.48

3　结论

针对口罩人脸识别问题,本文提出融合 PVT 多级特征的模型。将 PVT 的 SRA 替换为基于自注意力机制的 MSA 以提取更丰富的人脸特征,并通过特征融合使模型集中关注未被口罩遮挡的人脸区域。为了减少模型的参数量和运算量,本文提出了一种融合各 Stage 的 *cls* 向量的特征融合方法。最后,本文采用 Sub-center ArcFace 作为损失函数,适当放宽数据的类内约束来进一步提升模型的鲁棒性。实验表明,该模型能有效提取口罩人脸中的特征,在普通人脸、模拟口罩人脸和真实人脸数据集上均取得了出色的效果。相较于其他特征融合方法,本文所提方法具有参数量更小、精度更高的特点。

参考文献(References)

[1] 索继江,闫中强,刘运喜,等. 新型冠状病毒肺炎医院感染现状及预防控制策略与措施探讨[J]. 中华医院感染学杂志,2020,30(6):811-816.
SUO J J, YAN Z Q, LIU Y X, et al. Current status of hospital-acquired COVID-19 and strategies for prevention and control[J]. Chinese Journal of Nosocomiology, 2020,30(6):811-816.

[2] MANDAL B, OKEUKWU A, THEIS Y. Masked face recognition using ResNet-50[J]. arXiv preprint, 2021.

[3] 姜绍忠,姚克明,陈磊,等. 基于 CNN 与 Transformer 混合模型的口罩人脸识别方法[J]. 传感器与微系统,2023,42(1):144-148.
JIANG S Z, YAO K M, CHEN L, et al. Masked face recognition method based on CNN and Transformer hybrid model[J]. Transducer and Microsystem Technologies, 2023,42(1):144-148.

[4] LI Y D, GUO K, LU Y G, et al. Cropping and attention based approach for masked face recognition[J]. Applied Intelligence, 2021,51(5):3012-3025.

[5] QIAN H J, ZHANG P P, JI S J, et al. Improving representation consistency with pairwise loss for masked

face recognition[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops(ICCVW), 2021: 1462-1467.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017: 5998-6008.

[7] WANG W H, XIE E Z, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision(ICCV), 2021: 568-578.

[8] ZHANG J W, YAN X D, CHENG Z L, et al. A face recognition algorithm based on feature fusion[J]. Concurrency and Computation: Practice and Experience, 2022, 34(14): e5748.

[9] WANG H, WANG Y T, ZHOU Z, et al. Cosface: Large margin cosine loss for deep face recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 5265-5274.

[10] DENG J K, GUO J, YANG J, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4690-4699.

[11] DENG J K, GUO J, LIU T L, et al. Sub-center arcface: Boosting face recognition by large-scale noisy web faces[C]//European Conference on Computer Vision, 2020: 741-757.

[12] GUO Y D, ZHANG L, HU Y X, et al. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition[C]//European Conference on Computer Vision, 2016: 87-102.

[13] ANWAR A, RAYCHOWDHURY A. Masked face recognition for secure authentication[J]. arXiv: 2008.11104, 2020.

[14] HUANG G B, MATTAR M, BERG T L, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments[C]//Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, 2008.

[15] DENG W H, HU J N, ZHANG N H, et al. Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership[J]. Pattern Recognition, 2017, 66: 63-73.

[16] ZHENG T, DENG W. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments[R]. Beijing University of Posts and Telecommunications, Tech. Rep, 2018, 5(7).

[17] WANG Z Y, WANG G C, HUANG B J, et al. Masked face recognition dataset and application[J]. arXiv: 2003.09093, 2020.

[18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016: 770-778.

[19] HAN K, WANG Y H, TIAN Q, et al. Ghostnet: More features from cheap operations[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2020: 1577-1586.

[20] CHEN S, LIU Y, GAO X, et al. MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices[C]//Chinese Conference on Biometric Recognition, 2018: 428-438.

[21] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv: 2010.11929, 2020.

[22] DAI Y M, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//2021 IEEE Winter Conference on Applications of Computer Vision(WACV), 2021: 3560-3569.

[23] ZHANG D, ZHANG H, TANG J, et al. Feature pyramid transformer[C]//European Conference on Computer Vision, 2020: 323-339.

[24] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE Conference on Computer Vision, 2017: 618-626.

(责任编辑:郭芸婕)