基于多层次特征融合的 Transformer 人脸识别方法

夏桂书[1]，朱姿翰[1]，魏永超[2]，朱泓超[3]，徐未其[3]
(1. 中国民用航空飞行学院航空电子电气学院，德阳 618307；
2. 中国民用航空飞行学院科研处，德阳 618307；
3. 中国民用航空飞行学院民航安全工程学院，德阳 618307)

摘　要：卷积神经网络中的卷积操作只能捕获局部信息，而 Transformer 能保留更多的空间信息且能建立图像的长距离连接. 在视觉领域的应用中，Transformer 缺乏灵活的图像尺寸及特征尺度适应能力，通过利用层级式网络增强不同尺度建模的灵活性，且引入多尺度特征融合模块丰富特征信息. 本文提出了一种基于改进的 Swin Transformer 人脸模型——Swin Face 模型. Swin Face 以 Swin Transformer 为骨干网络，引入多层次特征融合模块，增强了模型对人脸的特征表达能力，并使用联合损失函数优化策略设计人脸识别分类器，实现人脸识别. 实验结果表明，与多种人脸识别方法相比，Swin Face 模型通过使用分级特征融合网络，在 LFW、CALFW、AgeDB-30、CFP 数据集上均取得最优的效果，验证了此模型具有良好的泛化性和鲁棒性.

关键词：人脸识别；Transformer；多尺度特征；特征融合
中图分类号：TP391　　**文献标志码**：A　　DOI：10.19907/j.0490-6756.2024.012002

Transformer face recognition method based on multi-level feature fusion

XIA Gui-Shu[1], *ZHU Zi-Han*[1], *WEI Yong-Chao*[2], *ZHU Hong-Chao*[3], *XU Wei-Qi*[3]
(1. Institute of Electronic and Electrical Engineering, Civil Aviation Flight University of China, Deyang 618307, China;
2. Department of Scientific Research Office, Civil Aviation Flight University of China, Deyang 618307, China;
3. College of Civil Aviation Safet Engineering, Civil Aviation Flight University of China, Deyang 618307, China)

Abstract: The convolutional operation in a convolutional neural network only captures local information, whereas the Transformer retains more spatial information and can create long-range connections of images. In the application of vision field, Transformer lacks flexible image size and feature scale adaptation capability. To solve this problems, the flexibility of modeling at different scales is enhanced by using hierarchical networks, and a multi-scale feature fusion module is introduced to enrich feature information. This paper propose an improved Swin Face model based on the Swin Transformer model. The model uses the Swin Transformer as the backbone network and a multi-level feature fusion model is introduced to enhance the feature representation capability of the Swin Face model for human faces. a joint loss function optimisation strategy is used to design a face recognition classifier to realize face recognition. The experimental results show that, compared with various face recognition methods, the Swin Face recognition method achieves best results on LFW, CALFW, AgeDB-30, and CFP datasets by using a hi-

收稿日期：2023-03-09
基金项目：西藏科技厅重点研发计划(XZ202101ZY0017G)；四川省科技厅重点研发项目(2022YFG0356)；中国民用航空飞行学院科研基金(J2020-126，J2020-040，J2021-056)
作者简介：夏桂书(1968－)，女，硕士，教授，研究方向为航空电子. E-mail：xgs19680922@163.com
通讯作者：朱姿翰. E-mail：1476514200@qq.com.

erarchical feature fusion network, and also has good generalization and robustness.

Keywords: Face recognition; Transformer; Multi-scale features; Feature fusion

1 引　言

目前,人脸识别算法主要分为两类:一种是基于手工特征的人脸识别算法,通过人工设计特征提取器提取人脸特征信息,再结合不同的分类算法实现人脸识别.另一种是基于深度学习的人脸识别算法[1],大部分算法是基于卷积神经网络(Convolution Neural Network,CNN)来实现.

基于手工特征的人脸识别方法可以分为基于几何特征[2]、基于模板匹配[3]和基于子空间的方法[4].基于几何结构特征的方法主要对人脸的几何特征点进行提取,以此来完成人脸识别;基于模板匹配的方法利用可变性模板对人脸面部特征进行抽取人脸特征向量,通过计算图像与模板特征向量之间的距离来判断人脸类别;基于子空间的方法将人脸高维数据映射到低维空间,通过 K-L 变换压缩技术来表示人脸特征[5].基于手工特征的人脸识别方法在受到光照变化、姿态变化等外在因素影响时,会造成人脸特征急剧变化,使得人脸识别准确度降低[6].

而基于深度学习的方法因其强大的特征提取能力逐渐取代了手工提取人脸特征的方法.2014 年,Facebook 团队提出的 DeepFace[7]方法使用 3D 模型将人脸对齐,再通过 CNN 来提取人脸特征,提高了面部识别的准确性;同年,港中文汤晓鸥团队提出 DeepID[8]在人脸识别过程中采用极大的分类准则,并把学习到的高级特征表达集合应用到人脸验证上,相比于传统的人脸识别算法泛化能力增强;2015 年 Google 团队提出的 FaceNet[9]通过深度学习结构将人脸特征映射到欧式空间中,利用三元组损失函数(Triplet Loss,TL)增大类间距离,缩小类内距离;2017 年,Jiang 等人[10]所提出的 RetinaNet 网络 One-stage 首次超越了 Two-stage 网络,Insightface 团队基于检测网络 RetinaNet 提出 Retinaface[11]网络,添加了 SSH 网络的三层级联检测模块,利用了特征金字塔等策略提升了人脸识别检测精度.

随着人工智能技术的快速发展,Transformer 在计算机视觉领域取得了重要的成果[12].目前,将 Transformer 应用于不同的计算机视觉任务,包括图像分类、目标检测以及视频处理等已经成为一个流行的趋势.例如,iGPT[13],BEIT[14]利用 Transformer 网络架构代替卷积神经网络完成图像分类;基于 Transformer 架构的 DETR[15]实现了端到端的目标检测.由于 CNN 中的卷积操作只能捕获局部信息,不能建立全局图像的长距离连接,而 Transformer 通过多头注意力操作能实现特征汇聚,增强其全局性.相比于 CNN,Transformer[16]保留更多的空间信息且能够捕捉到更多的特征信息.但由于视觉实体的大小差异很大,自然语言处理(Natural Language Processing,NLP)对象的大小是标准固定的,且图像中的像素与文本中的单词相比具有很高的分辨率.通过利用层级化 Transformer 可以增强网络在不同图像尺度下的建模能力.本文以 Swin Transformer[17]为骨干网络,设计了具有多层次特征融合的 Swin Face 人脸网络模型.其优点如下:(1) Swin Face 人脸模型采用了基于滑动窗口的多头注意力机制,有效建立了不同窗口之间的连接,有利于网络捕获图像的全局信息,提高网络的性能.(2) Swin Face 人脸模型引入了多层次特征融合方式,将各层级信息进行有效融合,获取到人脸模型的分层特征,提高了网络对人脸特征表达能力,弥补了使用单一特征在人脸特征提取上的不足.

2 Swin Face 模型

基于 Swin Face 的人脸识别网络模型整体采用层次化、多尺度的设计,包含三个主要模块,Patch Embed 模块、Swin Transformer 模块以及多尺度特征融合模块.整体网络结构如图 1 所示.

2.1 Patch Embed 模块

Patch Embed 模块包含了两个功能:块分割(Patch Partition)和块的线性嵌入(Linear Embedding),块分割负责将图片切成非重叠、等尺寸大小的块;线性嵌入层将每个块做降维采样,缩小图像分辨率.本文中人脸图像输入尺寸大小为 3×160×160,利用卷积操作将人脸映射成一组 Token,输出尺寸为 1×1600×196.具体流程图如图 2 所示.

2.2 Swin Transformer 模块

Swin Transfomer 模块主要由以下几部分构成:窗口多头自注意力层(Window Multi-head Self Attention, W-MSA)、滑动窗口多头自注意力层

(Shifted Window based Multi-head Self-attention，SW-MSA)、多层感知机(Multi Layer Perceptron，MLP)、标准化层(Layer Normalization，LN)[18]. Swin Transfomer 模块具体如图 3 所示.

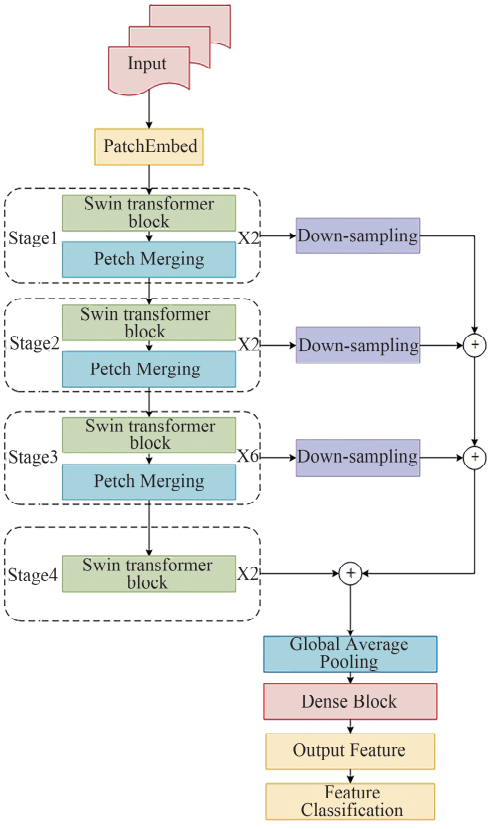

图 1　Swin Face 模型总体结构

Fig. 1　General structure of the Swin Face model

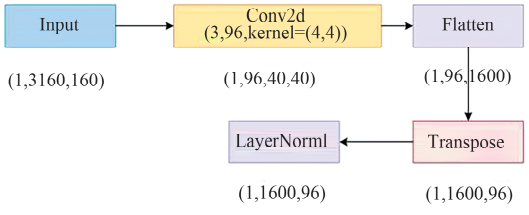

图 2　Patch Embed 模块流程图

Fig. 2　Patch Embed block flowchart

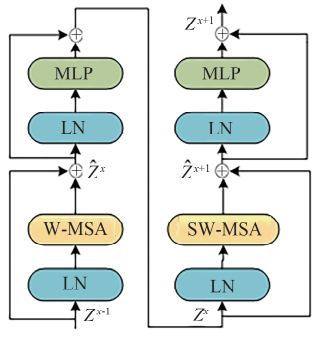

图 3　Swin Transformer 模块流程图

Fig. 3　Swin Transformer block flowchart

Swin Transformer 模块主要进行特征处理，每部分的输出如式(1)～式(4)所示.

$$\hat{Z}^X = W\text{-}MSA(LN(Z^{X-1})) + Z^{X-1} \tag{1}$$

$$Z^X = MLP(LN(\hat{Z}^X)) + \hat{Z}^X \tag{2}$$

$$\hat{Z}^{X+1} = SW\text{-}MSA(LN(Z^X)) + Z^X \tag{3}$$

$$Z^{X+1} = MLP(LN(\hat{Z}^{X+1})) + \hat{Z}^{X+1} \tag{4}$$

输入的特征图首先通过块分割层，将特征图划分为非重叠、等尺寸大小的块，通过线性嵌入层将块转换成一个长度为 96 的嵌入向量 Token，紧接着通过 Stage 阶段来进行处理：Stage1，Stage2，Stage3，Stage4. Stage1 中，Block 数量为 2，Tokens 的输出为 $\frac{H}{4}\times\frac{W}{4}$，输出维度为 C；Stage2 中，Block 数量为 2，Tokens 的输出为 $\frac{H}{8}\times\frac{W}{8}$，输出维度为 $2C$；Stage3 中，Block 数量为 6，Tokens 的输出为 $\frac{H}{16}\times\frac{W}{16}$，输出维度为 $4C$；Stage4 中，Block 数量为 2，Tokens 的输出为 $\frac{H}{32}\times\frac{W}{32}$，输出维度为 $8C$.

2.2.1　窗口多头自注意力　注意力机制是建立特征和特征之间的关系，以高权重聚焦重要信息，以低权重忽略不相关信息. 因此，引入注意力机制能有效增强空间编码能力，使得模型具有更强的鲁棒性与泛化性[19].

多头注意力机制是并行地从输入信息中选取多个信息，每个注意力关注输入信息的不同部分，然后再将每一组自注意力的结果拼接起来进行一次线性变换得到最终的输出结果. 多头自注意力机制首先将输入信息通过 Linear 操作生成 Q、K、V 三个权重向量，然后对每个头都进行自注意力操作，最后将每个头输出的结果进行 Concat 操作，通过 Linear 层输出最后的特性信息. 自注意力图如图 4 所示，多头注意力图如图 5 所示，多头注意力计算公式如式(5)和式(6)所示.

$$\text{Multihead}(Q,K,V) = \text{Concat}(head_1, head_2, \ldots, head_h)W^o \tag{5}$$

$$head = \text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK}{\sqrt{d}}\right)V \tag{6}$$

式中，Q 表示查询向量；K 表示键向量；V 表示值向量；$W^o \in R^{d\times d}$ 表示多头注意力权重矩阵；$head$ 表示多头注意力机制中头的个数；$Concat$ 表示拼接操作.

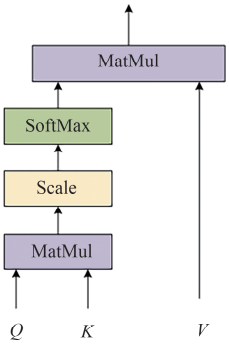

图 4　自注意力图
Fig. 4　Self-Attention chart

2.2.2　滑动窗口多头自注意力　窗口内分别计算自注意力,首先将 $H\times W\times C$ 的特征图划分为非重叠的窗口,窗口尺寸为 $L\times L$,窗口数量为 $\frac{H\times W}{L^2}$,然后在窗口内分别计算自注意力.

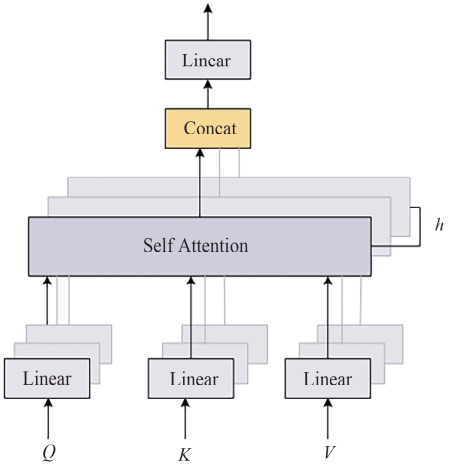

图 5　多头注意力图
Fig. 5　Multi-head attention chart

为实现跨窗口之间的信息融合与交互,使用向左上方循环移位的批处理计算方法,如图 6 所示.首先,对每个窗口内进行自注意力操作,如(1,2,3,4),(5,6,7,8),(9,10,11,12),(13,14,15,16),每个列表内的元素做自注意力运算,这样可以建立各自窗口之间的联系.然后,向左上方循环移位窗口,将图像补回原图像大小,在各自的窗口内再次做自注意力计算,可以建立(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16)之间的联系.通过移位之后,一个批处理窗口可能由几个在特征图中不相邻的子窗口组成,因此使用掩码机制将自注意力计算限制在每个子窗口内.通过循环移位,批处理窗口的数量保持与常规窗口分区相同.

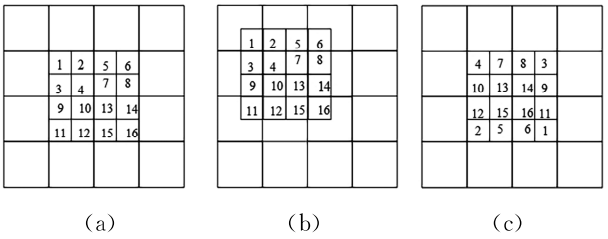

图 6　左上方循环移位计算流程图
Fig. 6　Top left circular shift calculation flowchart

2.3　多尺度特征融合

通常来说,卷积神经网络在进行特征提取时,获取到的低级特征具有丰富的几何信息,但是语义信息表征能力较弱,而获取到的高级特征具有丰富的语义信息,但是缺乏空间细节特征[20].通过多尺度的融合方式,利用不同尺度的卷积核对目标特征进行提取,将网络的低级信息与高级信息进行特征融合,有效提高特征的丰富度,能够增强网络的表征能力.因此,多尺度特征融合不仅能够减少不同特征通道层之间的语义差距,提高网络的表征能力,而且能够丰富特征的结构信息.具体的多尺度特征融合流程图如图 7 所示.

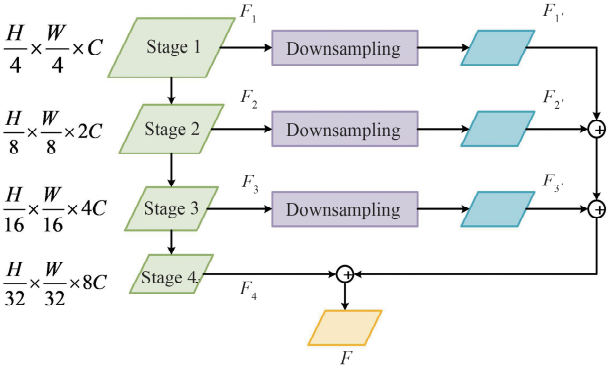

图 7　多尺度特征融合流程图
Fig. 7　Multi-scale feature fusion flowchart

为了增强特征尺度适应能力,从 4 个 Stage 阶段分别抽取的特征 F_1、F_2、F_3 和 F_4,经过 Downsampling 操作分别得到特征 F_1'、F_2' 和 F_3',最后将特征 F_1'、F_2'、F_3' 和 F_4 进行求和操作得到融合特征 F.计算公式如式(7)~式(10)所示.

$$F_1'=\text{Downsampling}(F_1) \tag{7}$$

$$F_2'=\text{Downsampling}(F_2) \tag{8}$$

$$F_3'=\text{Downsampling}(F_3) \tag{9}$$

$$F=\text{Add}(F_1',F_2',F_3',F_4) \tag{10}$$

式中,Downsampling 表示下采样操作;Add 表示求和操作.

3 实　验

3.1 数据集

本文所使用的数据集如表 1 所示. 在人脸领域，通常选用的训练数据集为 CASIA-WebFace，其适用于非约束环境下人脸识别科学研究. 实验所使用的测试数据集为 LFW、CALFW、CPLFW、AgeDB-30[21]、CFP[22]. LFW 人脸数据集是目前人脸识别的常用数据集，其中所提供的人脸图片均来自于不同的自然场景，包括不同姿态、光照、表情等异质人脸图像；CALFW 是基于 LFW 数据集标注的跨年龄数据集；CPLFW 是基于 LFW 数据集标注的跨姿态数据集；AgeDB-30 数据集包括不同姿态、表情、年龄、性别的图片；CFP 数据集中每个人都有 10 张正面图像和 4 张侧面图像. 本文从每个测试数据集中随机选取 6000 对人脸组成人脸识别图像对，其中 3000 对属于同一个人的两张图像，3000 对属于不同的两个人脸图像. 表 1 中，IDs 表示身份数量；Imgs 表示图片数量.

表 1　实验相关数据集

Tab. 1　Experiment-related datasets

	Dataset	IDs	Imgs
训练集	CASIA-WebFace	10575	494 414
测试集	LFW	5749	13 233
测试集	CALFW	5749	13 233
测试集	CPLFW	5749	13 233
测试集	AgeDB-30	440	12 240
测试集	CFP	500	7000

3.2 数据处理

人脸图像根据采集环境的不同及环境的干扰，如光照变化、遮挡、距离远近等，需要对图像进行预处理以保证人脸图像的质量. 因此，为了更好地对人脸进行特征提取，需对人脸图像进行检测. 本文使用 DNN 模型对人脸进行检测，具体的人脸检测过程如图 8 所示.

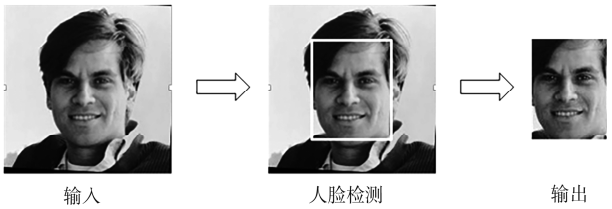

图 8　人脸检测过程

Fig. 8　Face detection process

3.3 实验环境

本文所采用的实验环境为：操作系统 Windows 10 CPU Intel（R） Core（TM） i9-12900K CPU @ 3.20 GH，NVIDIA GeForce RTX 3090Ti. 深度学习框架为 Pytorch，CUDA 版本为 11.6.

在验证算法性能时，所采用的参数设置保持一致. 设置 Batch Size 为 128，设置训练迭代次数为最大为 50；使用 Adam 优化网络模型，初始学习率为 0.001；损失函数设置，采用三元组损失函数和交叉熵损失函数，三元组损失函数的阈值设置为 0.4.

3.4 损失函数

为了提高网络的表征能力，将多种损失函数进行联合，达到联合优化的效果. 本文采用的是将三元组损失函数和交叉熵损失函数（Cross Entropy Loss，CE）相结合的策略进行训练的方法[23].

3.4.1　三元组损失　三元组损失函数需要从训练样本中选取目标样本、正样本、负样本，利用样本之间的距离作为约束，增大不同类样本的距离，缩小同类样本的距离[24]. 相比其他分类损失函数，Triplet Loss 通常能在训练中学习到更好的细微的特征，更特别的是 Triplet Loss 能够根据模型训练的需要设定一定的阈值 mgn，设计者可以通过改变 mgn 的值来控制正负样本的距离. Triplet Loss 损失函数为

$$L_{TL}=\max(d(a,p)-d(a,n)+mgn,0) \tag{11}$$

式中，a 表示目标图像；p 表示正样本，与 a 是同一类别样本；n 表示负样本，与 a 是不同类别的样本；mgn 表示阈值，是一个大于 0 的常数.

3.4.2　交叉熵损失函数　本文在分类问题中，采用的是交叉熵损失函数，此损失函数擅长学习类间信息，能够增强网络的性能. 交叉熵损失函数为

$$L_{CE}=-\sum_{i=1}^{N} y_i\hat{y}_i \tag{12}$$

式中，y_i 为样本标签；$\hat{y}_i$ 为网络的输出值.

3.4.3　联合损失函数　本文将三元组损失函数和交叉熵损失函数联合起来作为人脸识别网络训练所使用的损失函数，使得网络在联合函数优化下，提高网络的表征能力. 联合损失函数为

$$L=L_{TL}+L_{CE} \tag{13}$$

3.5 评估指标

本文方法在 LFW、CALFW、CPLFW、AgeDB-30、CFP 等 5 个数据集上进行评估，以准确率作为评价指标. 准确率计算公式为

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \tag{14}$$

其中,TP 为真阳性数;TN 为真阴性数;FP 为假阳性数;FN 为假阳性数.

3.6　实验结果及分析

3.6.1　与不同方法实验对比　为验证本文算法的泛化性及可行性,与 DeepID[8]、Retinaface[11]、Swin Transforme[17]、EdgeNext[25]、CMT[26] 五种骨干网络进行实验对比,以平均准确率为度量,实验结果如表 2 所示. 表 2 中,Swin Transformer 记为 Swin T.

表 2　不同方法准确率定量对比

Tab. 2　Quantitative comparison of the accuracy of different methods

方法	LFW /%	CALFW /%	CPLFW /%	AgeDB-30 /%	CFP/%
DeepID	87.60	73.23	65.05	68.83	66.13
Retinaface	96.18	88.00	79.70	81.18	95.02
EdgeNext	94.68	84.67	75.26	78.63	81.85
CMT	96.72	88.58	81.44	80.92	91.03
Swin T	96.60	87.98	80.07	81.34	95.09
Swin Face	97.10	88.72	80.97	82.19	95.50

与其他 5 种人脸识别算法相比,在人脸姿态变化、年龄变化、光照变化等情况下,本文所提出的 Swin Face 算法在其中的四个基准测试数据集上均取得最优的效果,在 CPLFW 数据集上取得的效果与最优相差 0.47%. 从总体来看,Swin Face 模型对同一身份的人脸变化具有良好的鲁棒性和泛化性.

与 Swin Transformer 模型相比,Swin Face 模型通过融合分级特征的方式,建立浅层信息与深层信息之间的关联性,提高了网络对人脸的特征表达能力. 人脸识别精度提升了 0.41%～0.9%. 由此可得,多层次特征融合模块,充分利用了低级语义信息和高级语义信息,实现人脸特征聚集,由此说明了融合多层次特征模块提升了对人脸识别模型的准确率.

3.6.2　消融实验　为探究多层次特征融合模块的性能,本文从两个方面进行对比实验.(1) 模型性能评估,不同方法在统一的硬件设施下进行运算,具体结果如表 3 所示.(2) 泛化性评估. 两种方法在基准测试数据集进行验证测试,参数量与速度对比结果如图 9 和图 10 所示.

表 3　参数量与速度对比

Tab. 3　Number of participants vs. speed

Model	Params/M	FLOPs/G	FPS
Swin Transformer	27.59	2.18	29.9
Swin Face	28.04	2.27	29.1

表 3 中,Params 表示模型训练参数;FLOPs 表示浮点运算数;FPS 表示模型推理速度. 由表 3 可知,在模型训练阶段,相比于 Swin Transformer 模型,Swin Face 模型参数量增加了 0.45 M,FLOPs 提高了 0.09 G,可知 Swin Face 训练速度有所降低. 在模型推理阶段,Swin Face 推理速度略低于 Swin Transformer. 分析可知,由于 Swin Face 融合了多层次特征网络,增加了网络的复杂度,导致模型整体速度下降. 通过对图 9 和图 10 平均准确率曲线分析可知,即使是不同基准测试数据集,但随着训练进程的继续,Swin Face 模型相较于 Swin Transformer 模型曲线波动更小,更加平稳. 相比于 Swin Transformer 模型,Swin Face 模型在 LFW、CALFW、CPLFW、AgeDB-30、CFP 上提高的人脸识别的准确率分别为 0.50%、0.74%、0.90%、0.85%、0.41%,验证了 Swin Face 模型的可行性. 从模型性能可知,虽然在模型训练阶段增加了训练参数及模型复杂度,但其对模型的整体影响较小,且人脸识别精度提升最高达到 0.90%,验证了融合多层次特征模块对人脸识别模型的有效性.

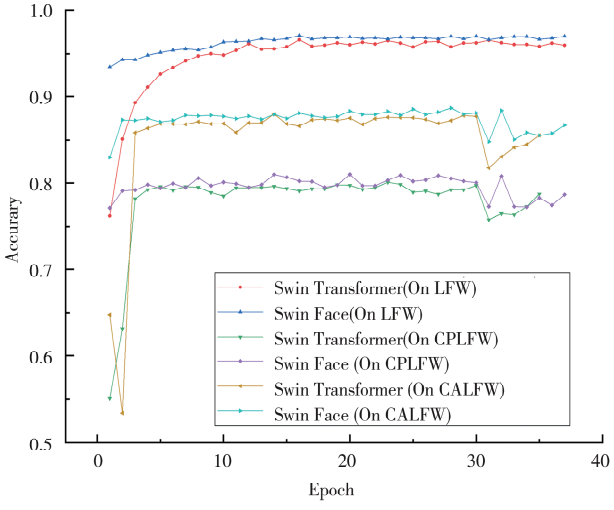

图 9　在 LFW、CPLFW 和 CALFW 上的准确率曲线

Fig. 9　Accuracy curves on LFW, CPLFW and CALFW

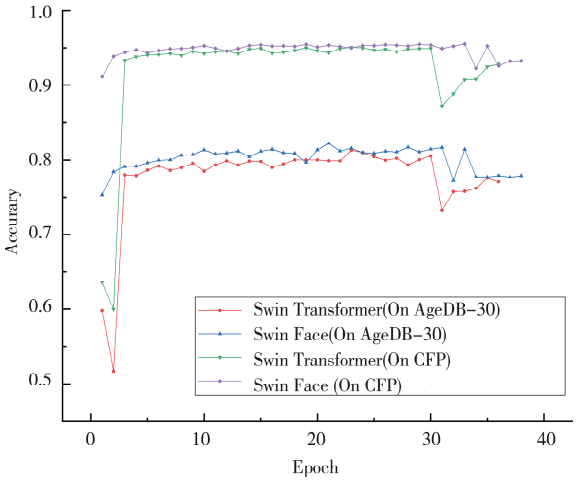

图 10　在 AgeDB-30 和 CFP 数据集上的准确率曲线
Fig. 10　Accuracy curves on AgeDB-30 and CFP

4　结　论

本文提出 Swin Face 人脸识别模型，通过引入多层次特征融合模块，将浅层信息与深层信息进行有效融合，再结合 Swin Transformer 的多头注意力机制，获取全局依赖关系，构建层次映射，提高网络的全局建模能力. 在训练过程中，此模型使用联合损失函数和 Adam 优化策略，增强了特征间的约束，进一步提高了网络泛化能力. 实验结果表明，该模型在不同数据集上均取得最优的效果，说明了多层次特征融合模型具有良好的鲁棒性. 但是本文方法仍存在模型参数量大和计算复杂度高的问题，因此，后续的研究工作中，应考虑优化算法使模型收敛速度加快，降低模型的复杂度.

参考文献：

[1] Kang L, Yan T. An Analysis of Face Recognition Algorithms Based on Deep Learning[J]. Yangtze River Inform Comm, 2022, 35: 83.[康磊, 闫涛. 基于深度学习的人脸识别算法浅析[J]. 长江信息通信, 2022, 35: 83.]

[2] Wu K, Zhou M Y, LI G Y, *et al*. Face expression recognition based on angular geometric features [J]. Comp Appl Softw, 2020, 37: 120.[吴珂, 周梦莹, 李高阳, 等. 基于角度几何特征的人脸表情识别[J]. 计算机应用与软件, 2020, 37: 120.]

[3] Du W C, Dang Z P, Zhao Q J, *et al*. Face alignment based on local shape constraint networks[J]. J Sichuan Univ(Nat Sci Edi), 2017, 54: 953.[杜文超, 邓宗平, 赵启军, 等. 基于局部形状约束网络的人脸对齐[J]. 四川大学学报(自然科学版), 2017, 54: 953.]

[4] Lv F F. Analysis and research on face recognition method based on subspace [J]. Comp Knowl Technol, 2020, 16: 185.[吕芳芳. 基于子空间的人脸识别方法的分析与研究[J]. 电脑知识与技术, 2020, 16: 185.]

[5] Wang Y H. Face recognition based on K-L transform and singular value decomposition [J]. J Hebei Inst Water Resour Elect Power, 2020, 30: 38.[王银花. 基于 K-L 变换和奇异值分解的人脸识别[J]. 河北水利电力学院学报, 2020, 30: 38.]

[6] Li X F, You Z S. Large-pose face recognition based on 3D-2D mapping [J]. J Sichuan Univ(Nat Sci Ed), 2022, 59: 61.[李晓峰, 游志胜. 基于 3D-2D 映射的大姿态人脸识别[J]. 四川大学学报(自然科学版), 2022, 59: 042003.]

[7] Wang M, Deng W. Deep face recognition: a survey [J]. Neurocomputing, 2021, 429: 215.

[8] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10000 classes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014.

[9] Schroff F, Kalenichenko D, Philbin J. Facenet: a unified embedding for face recognition and clustering [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles: IEEE Computer Society, 2015.

[10] Jiang C, Ma H, Li L. IRNet: an improved retinanet model for face detection [C]//Proceedings of the 7th International Conference on Image, Vision and Computing (ICIVC). Los Angeles: IEEE Computer Society, 2022.

[11] Deng J, Guo J, Ververas E, *et al*. Retinaface: single-shot multi-level face localisation in the wild [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020.

[12] Khan S, Naseer M, Hayat M, *et al*. Transformers in vision: a survey [J]. ACM Comput Surv, 2022, 54: 1.

[13] Chen M, Radford A, Child R, *et al*. Generative pretraining from pixels [C]//Proceedings of the International Conference on Machine Learning. Vienna: [s. n.], 2020.

[14] Devlin J, Chang M W, Lee K, *et al*. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of Annual Conference of the North American Chapter of the

Association for Computational Linguistics. Minneapolis: [s. n.], 2019.

[15] Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Berlin: Springer, 2020.

[16] Lin T, Wang Y, Liu X, *et al*. A survey of transformers [J]. Artif Intell Rev, 2022, 2: 04554.

[17] Liu Z, Lin Y, Cao Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows [C]//Proceedings of the IEEE International Conference on Computer Vision. Los Angeles: IEEE Computer Society, 2021.

[18] Liu W T, Lu X M. Research progress of Transformer based on computer vision [J]. Comp Eng Appl, 2022, 58: 1.[刘文婷, 卢新明. 基于计算机视觉的 Transformer 研究进展[J]. 计算机工程与应用, 2022, 58: 1.]

[19] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning [J]. Neurocomputing, 2021, 452: 48.

[20] Xia H, Ma J, Ou J, *et al*. Pedestrian detection algorithm based on multi-scale feature extraction and attention feature fusion [J]. Digit Signal Process, 2022, 108: 103311.

[21] Moschoglou S, Papaioannou A, Sagonas C, *et al*. Agedb: the first manually collected, in-the-wild age database [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington: IEEE, 2017.

[22] Sengupta S, Chen J C, Castillo C, *et al*. Frontal to profile face verification in the wild [C]//Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Los Angeles: IEEE, 2016.

[23] Rybicka M, Kowalczyk K. On parameter adaptation in softmax-based cross-entropy loss for improved convergence speed and accuracy in dnn-based speaker recognition[C]//Interspeech. Shanghai: ISCA, 2020.

[24] Zhang X Y, You M G, Zhu J, *et al*. Face recognition based on joint loss function for small-scale data [J]. J Beijing Inst Technol, 2020, 40: 163.[张欣彧, 尤鸣宇, 朱江,等. 基于联合损失函数的小规模数据人脸识别[J]. 北京理工大学学报, 2020, 40: 163.]

[25] Maaz M, Shaker A, Cholakkal H, *et al*. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications [C]// European Conference on Computer Vision. Berlin: Springer, 2023.

[26] Guo J, Han K, Wu H, *et al*. Cmt: convolutional neural networks meet vision transformers [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022.