Computer Science and Application 计算机科学与应用, 2023, 13(3), 301-310
Published Online March 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.133029

Hans汉斯

HSANet：混合型自我注意力网络识别微整容人脸方法

帕孜来提·努尔买提，古丽娜孜·艾力木江*

伊犁师范大学网络安全与信息技术学院，新疆 伊宁

收稿日期：2023年2月5日；录用日期：2023年3月3日；发布日期：2023年3月14日

摘　要

微整容给在日常生产中给人脸识别技术带来了新的挑战，因人脸特征变化较大导致对原人脸正确识别率较低，针对现象，该实验提出了一种混合型自我注意力块结构，用于识别面部特征变化的人脸，为此自制了26类微整容小样本图片数据集。将自我注意力融合到残差网络的瓶颈块中，提高了混合型自我注意力块对图片各区域特征的捕获能力，在对小样本微整容数据集的实验表明，该实验提出的混合型自我注意力网络有较高的正确识别率：89.70%，相比ResNet50正确识别率提高了2.65%，改进连接的混合型自我注意力模型比未改进连接的混合型自我注意力模型正确识别率提高了1.12%，网络性能也有所提升。

关键词

卷积神经网络，残差网络，瓶颈块，自我注意力，混合型自我注意力网络

HSANet: Hybrid Self-Attention Network Recognition Facial Micro Plastic Method

Pazilaiti Nuermaiti, Gulinazi Ailimujiang*

School of Network Security and Information Technology, Yili Normal University, Yining Xinjiang

Received: Feb. 5th, 2023; accepted: Mar. 3rd, 2023; published: Mar. 14th, 2023

Abstract

Due to the large changes in facial features, the correct recognition rate of the original face is low. In view of the phenomenon, this experiment proposed a hybrid self-attention block structure for recognizing faces with facial features changes. For this reason, 26 kinds of micro-plastic surgery

*通讯作者。

文章引用: 帕孜来提·努尔买提, 古丽娜孜·艾力木江. HSANet: 混合型自我注意力网络识别微整容人脸方法[J]. 计算机科学与应用, 2023, 13(3): 301-310. DOI: 10.12677/csa.2023.133029

small sample image data sets were made by ourselves. Integrating self-attention into the bottleneck block of the residual network improves the ability of the hybrid self-attention block to capture the features of each region of the image. The experiment on the small sample micro-plastic data sets shows that the hybrid self-attention network proposed in this experiment has a higher correct recognition rate: 89.70%, the correct recognition rate increased by 2.65% compared with ResNet50, and the correct recognition rate of the hybrid selfattention model with improved connection increased by 1.12% compared with the hybrid self-attention model without improved connection, and the network performance was also improved.

Keywords

Convolutional Neural Networks, Residual Networks, Bottleneck Blocks, Self-Attention, Hybrid Self-Attention Network

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

人脸识别技术在国防安全、视频监控、逃犯追踪和身份认证等方面发挥着重要作用。近年来，“深度学习”一词蔓延到了众多领域，在人工智能领域(Artificial Intelligence, AI)中，我们理解的深度学习(Deep Learning, DL)是类似人脑结构的多层网络结构，其学习过程也高度模仿了人脑对事物认知的基本过程，通过某种指定方式的训练后，会像人脑一样对指定的事物进行计算、分辨判断和自我优化。深度学习在图像分类、目标检测和实例分割等任务中学习和优化[1]，在大量深度学习众多图片识别的网络架构里卷积神经网络[2] (Convolutional Neural Network, CNN)是优选之一。在 CNN 架构中随着网络层的加深可以提取更多易于识别的特征信息，对普通网络(Plain Network)而言，随着 CNN 网络层数的加深，难以用优化的算法训练，并且其特有的网络性能也会逐步退化，其根本原因就是随网络层的加深，梯度也会逐步消失。为此，何凯明队在普通网络的基础上设计了残差网络[3] (ResNets)，在深层网络中加入残差设计，可以通过对残差块(Residual Block) [4]做计算来优化网络性能，梯度消失问题得到了解决。

ResNets 在过去几年是深度学习的首先网络，但是随着数据量的增加，ResNets 也面临着通过加深网络层数来提高模型鲁棒性的挑战。近几年，随着自我注意力机制(Self-Attention) [5]在自然语言处理(Natural Language Processing, NLP) [6]领域中愈加火热，研究者们最终将 Self-Attention 的优势运用到计算机视觉(Computer Vision, CV) [7]，研究出了远近交互性较强的一些网络模型(有：纯注意力模型的 SANet [8]和 Axial-SASA [9]，早期他们提出 Self-Attention 可以作为卷积模块的增强，以及另一种方向混合注意力模型的 AA-ResNet [10]和 BotNet [11]是将 CNN 与 Self-attention 结合在单个块内)，本文提出了一个新观点也属于混合注意力网络模型(Hybrid Self-Attention Net, HSANet)：用 Self-Attention 替换 ResNet50 Blotteneck 块的 *conv*3×3，并将块内 *conv*3×3 放在了 Identity 连接上，再把 Relu 激活函数移到快外，重造了结构类似于 Blotteneck 块的混合型 Self-Attention 块。

2. 相关知识

2.1. CNN 存在问题

适于监督学习的大多数 CNN 模型，面临着海量数据和捕获卷积长距离交互挑战，最初对于这些困难，

CNNs 是通过加深网络层数，来获得更细微的特征。CNNs 做卷积操作时，卷积后的每个独立像素之间与其边围的像素存在着一定的关联度，用于卷积操作的过滤器(Filter)只能提取局部信息，通过更多层的层卷积操作后才能将图像各区域的特征关联起来。

2.2. 残差结构介绍

残差块(Residual Block)是加深残差网络层的关键，快捷连接(Shortcut Connections) [12]和恒等映射(Identity Mapping) [13]是残差块的灵魂，其中恒等映射是加深网络层的核心，残差结构至少由两个残差块堆叠而成。为节省计算复杂度，将原先残差结构，如图 1 的两个 3 × 3 的卷积层，用图 2 的1×1+3×3+1×1的卷积层替换，其中两个 *conv*1×1 分别完成维度压缩和还原维度的工作，因而避免了梯度的消失，让深层网络的性能得到了质的飞跃。如图 2 所示的结构，称其为瓶颈(Bottleneck)结构，这种残差结构实现了通道数的先减少再增加。

恒等映射最终目标是尽量让输出与输入相近[14]，即公式(1)所示是恒等映射工作原理，公式(2)的 3 个分式所示是对恒等映射的最终目标的分步解释。

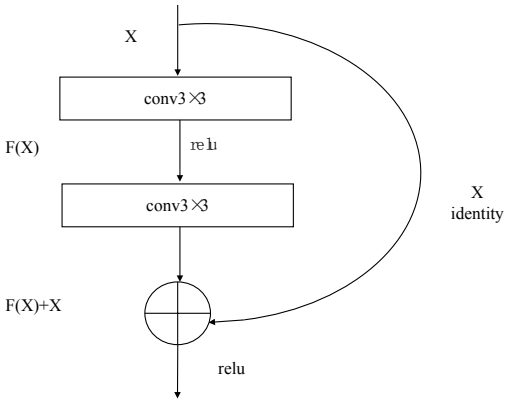

Figure 1. Residual structure
图 1. 残差结构

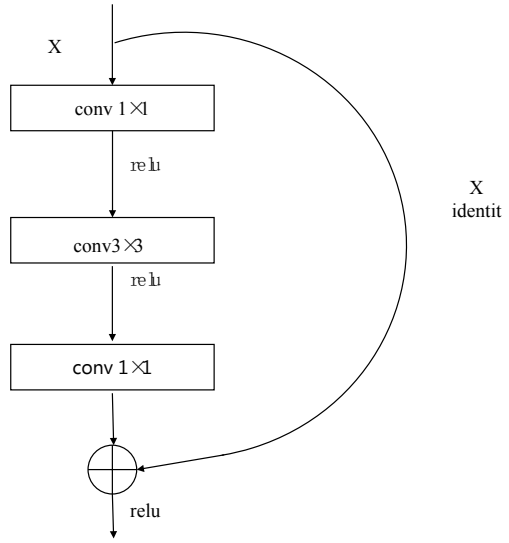

Figure 2. Residual bottleneck structure
图 2. 残差瓶颈结构

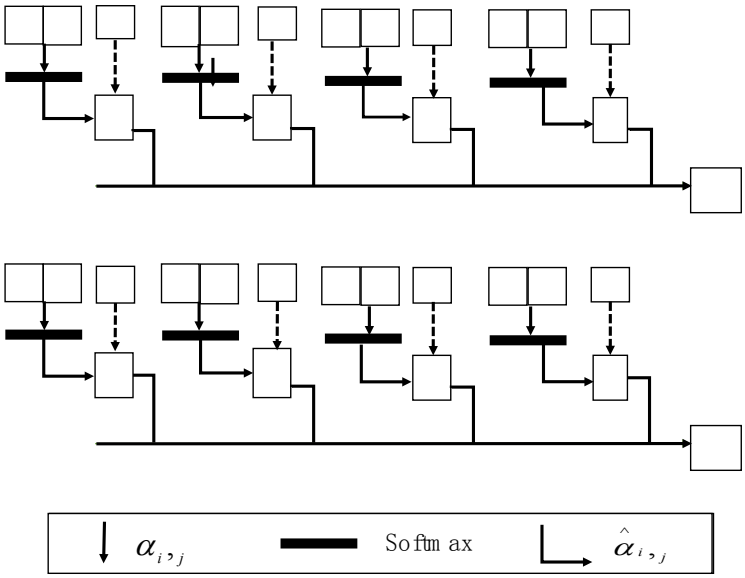

Figure 3. Self-attention mechanism workflow summary
图 3. 自我注意力机制工作流程概括

$$F(x)+x=H(x) \tag{1}$$

$$H(x)-x=F(x) \tag{2-1}$$

$$F(x)=0 \tag{2-2}$$

$$H(x)=x \tag{2-3}$$

在上式(1)，(2-1)至(2-3)中 x 表示输入，$F(x)$ 表示经过卷积操作后的输出也是下一个卷积的输入，$F(x)+x=H(x)$ 是最终的输出，上文提过恒等映射是恒是 $H(x)=x$，实际上我们只能逼近恒等映射，在学习过程参数量会有所消耗，因此最终目的不是让 $F(x)=0$，而让其逼近零。

2.3. 注意力机制

相比CNNs的短距离捕获信息，模拟人类注意力的注意力机制模型，可以更高效的捕捉长距离信息之间的关联度[15]。在深度学习里，注意力机制旨在从全局信息中筛选更具特色的信息，从而过滤无用信息。如公式(3)~(8)是自我注意力机制得主要工作原理。如图 3 所示是自我注意力机制工作流程概括。

$$q^i=w^q a^i \tag{3}$$

$$k^i=w^k a^i \tag{4}$$

$$v^i=w^v a^i \tag{5}$$

$$\alpha_{1,i}=\frac{q^1 k^i}{\sqrt{d}} \tag{6-1}$$

$$\alpha_{1,j}=\frac{q^1 k^j}{\sqrt{d}} \tag{6-2}$$

$$\alpha'_{i,j}=\frac{\alpha_{1,i}}{\sum_j \alpha_{1,j}} \tag{6-3}$$

$$(Q,K,V)=\text{Softmax}\frac{QK^{\mathrm{T}}}{\sqrt{d^{k}}}V \tag{7}$$

$$b^{n}=\sum_{i}\alpha'_{1,i}v^{i} \tag{8}$$

Self-Attention 是基于查询向量 Query (可用 Q 或 q 表示)、键向量 Key (可用 K 或 k 表示)和值 Value 向量(可用 V 或 v 表示)三个变量计算像素间的关联度。公式(3)是输入向量 α^{i} 与变换矩阵 w^{q} 相乘得到 q^{i}，(4)和(5)与式(3)同理。通过公式(6)计算和之间点乘计算相关性 α，并除以 $\sqrt{d}$ 来平衡梯度($\sqrt{d}$ 是缩放因子)，经过公式(7) Softmax 归一化操作后的 α 最终与 v 点乘，得到加权和的表示，依次累计操作得到公式(8)的作为位置关联度的 b^{n}。

3. 实验

3.1. 数据准备

用 python 对内娱网进行合法爬虫，批量下载明星图片(共 26 位明星，其中男 12 人，女 14 人，每人有 10 多张，共 283 张，还有一些在整容网上已公开的人脸图片共 378 张)。设置尺寸为 224 px × 224 px，确保每张人脸图像是彩色(通道为 RGB)的，对收集好的图片进行数据清洗，每个类以人名命名，最终生成较干净的数据集。因该实验数据集较小，将生成的数据集图片按 8:2 划分为训练集和测验集，通过训练得到的特征信息保存至 Log 文件里。

超参设置：使用交叉熵损失函数，让学习率衰减方式初始为 lr = 0.005，alpha = 0.25，expansion = 4，并以 batch_size 为 4 训练 50 个 epoch。通过数据增强对数据集图片进行随机裁剪、随机左右翻转和通过对图片进行遮挡，降低特征间的关联度，加强微整容特征变化程度。以防模型过拟合。

针对小样本微整容数据集的块内混合注意力模型 HSANet 受 Bottleneck Transformer 启发，从以下两方面做出了改进：1) 将 ResNet50 Bottleneck 中的 3 × 3 卷积替换成 Self-Attention；2) Identity 连接上的 x 替换成 3 × 3 卷积。如图 4 所示的是改进后的混合型自注意力单元块工作原理。

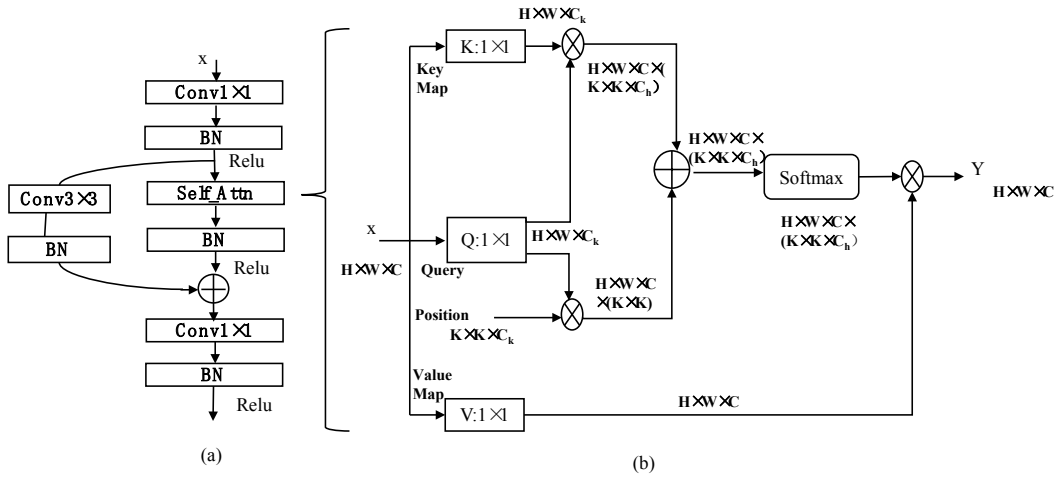

Figure 4. Working principle of hybrid self-attention block
图 4. 混合型自注意力单元块工作原理

首先在图 4 左半(a)部分，输入在线性块内 *conv*1×1 后，其输出进入 Self-Attention 中，同时 *conv*1×1 后的输出进入 Identity 连接的 *conv*3×3，Identity 连接跨 Self-Attention 层后，与线性残差块内 Self-Attention

的输出进行和操作后，结果再输送到 *conv*1×1 卷积，对最后的输出进行 Relu 激活函数。如图 4 右半(b)部分，其原理和图 4 所述相差无几。经过线性块内 *conv*1×1 后得输出特征图 x：$H\times W\times C$，$H\times W$ 是卷积后压缩得到的图，C 是通道，x 进入 Self-Attention 机制，将 k 与 q 相乘可得局部相关矩阵 $H\times W\times C\times\left(K\times K\times C_h\right)$ 记为 a'，C_h 为头部编号，每个 $K\times K$ 网格的相对位置 Position(P)与 q 矩阵相乘结果是每个像素的位置信息，接下来 a' 与每个像素的位置信息进行和操作后，在每个 C_h 通道维度上进行 Softmax 操作得注意力矩阵 A，是重塑每个像素的空间信息为 C_h 个局部注意力矩阵，A 与 V 聚合后最终得 Y。在 CV 中 Self-Attention 关注像素位置之间的关联性。

在 Identity 连接上设置 *conv*3×3 的优势在于，Identity 连接上 *conv*3×3 与 BN 可以维 Bottleneck 块结构。用 Self-Attention 替换线性网络中残差块内 *conv*3×3，等同于用一个向量组完成多个一维向量的工作，降低对低关联特征的关注，因此会提高对高关联度信息的注意。

该实验模型通过改进 ResNet50 直筒网络的残差块差块和连接，在残差块内用自我注意力块替换 *conv*3×3 的 Identity 连接，得到新的混合型自我注意力块为 HSANet。

3.2. 实验结果

为了验证混合型自我注意力方法在小样本微整容数据集上的识别效果，使用现有网络模型 ResNet50 和改进后的网络模型 HSANet(x)以及 HSANet (3 × 3) (两种对比混合模型 Identity 连接上的参数设置分别是 x 和 *conv*3×3) 3 种网络模型进行了对比，用于训练 3 种网络的超参数设置一致且训练过程相同。如图 5~8 所示小样本微整容数据集上分别在 ResNet50、HSANet(x)和 HSANet (3 × 3)上的识别效果。

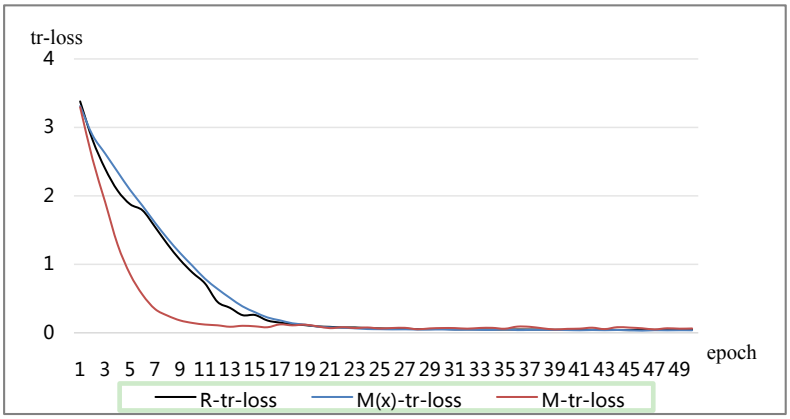

Figure 5. Loss comparison diagram of training set
图 5. 训练集的损失对比图

图 5 中纵坐标 tr-loss 表示训练集损失，横坐标 epoch，图中曲线 R-tr-loss 表示 ResNet50 的训练损失，Mx-tr-loss 表示 Identity 连接是 x 的自我注意力混合模型的训练损失，M-tr-loss 表示 Identity 连接是 3 × 3 的自我注意力混合模型的训练损失。从图 6 中的前 20 个 epoch 可以看到 M-tr-loss 曲线的损失下降快，后 30 个 epoch 中 M-tr-loss 的损失存在较小的梯度。

图 6 中纵坐标 tr-acc 表示训练集正确识别率，横坐标是 epoch，图中曲线 R-tr-acc 表示 ResNet50 的正确识别率，Mx-tr-acc 表示 Identity 连接是 x 的自我注意力混合模型训练正确识别率，M-tr-acc 表示 Identity 连接是 3 × 3 的自我注意力混合模型训练正确识别率。在图中前 20 个 epoch 可以发现混合模型 HSANet (3 × 3)和 HSANet (x)正确识别率(acc)上升速度快并且正确识别率高，其中 M-tr-acc (即 Identity 连接是 3 × 3 的自我注意力混合模型)的正确识别率稳居最高。三个网络模型的平均正确识别率(Avg acc)和最高正确

识别率分(Max acc)为：R-tr-acc 的 Avg acc 是 89.0350%，Max acc 是 98.1418% ，Mx-tr-acc 的 Avg acc 是 91.7801%，Max acc 是 98.7654%，M-tr-acc Avg acc 是 94.4301%，Max acc 是 98.7757%。

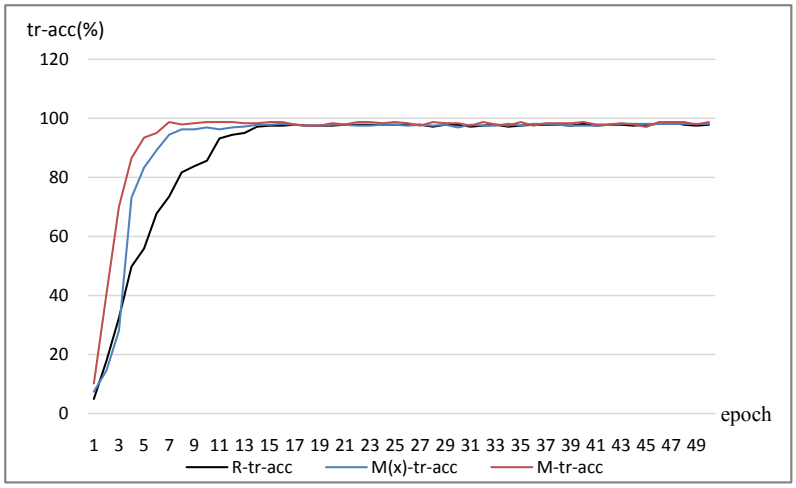

Figure 6. Comparison graph of correct recognition rate of training set
图 6. 训练集正确识别率对比图

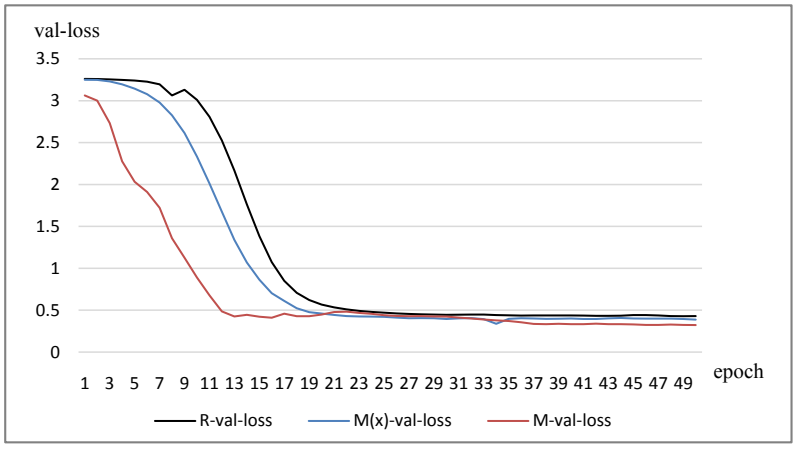

Figure 7. Loss function comparison diagram of verification set
图 7. 验证集的损失函数对比图

图 7 中纵坐标是 val-loss 表示验证集损失，横坐标是 epoch，在其前 20 个 epoch 可以发现混合模型(Mx-val-loss 和 M-val-loss 的混合型自我注意力模型)损失函数下降速度快，在后 10 个 epoch 中相对于其他 2 个损失函数，M-val-loss (即 Identity 连接是 3 × 3 的自我注意力混合模型)损失是最低的。

图 8 中纵坐标是 val-acc 表示验证集正确识别率，横坐标是 epoch，其前 12 个 epoch 可以发现混合模型(Mx-val-loss 和 M-val-loss 的混合型自我注意力模型)正确识别率(acc)上升速度快并且正确识别率高，其中 M-val-loss (即 Identity 连接是 3 × 3 的自我注意力混合模型)的正确识别率上升速度最快且最高。R-val-acc 的 Avg acc 是 66.8275%，Max acc 是 87.0270%，M-val-acc 的 Avg acc 是 72.1407%，Max acc 是 88.6486%，Mx-val-acc 的 Avg acc 是 76.0003%，Max acc 是 89.7011%。从图 6~8，可以发现 HSANet (3 × 3)的损失和正确识别率是 3 种网络中最理想的。

用模型评估指标 f1-score、macor avg 和 weight avg，在小样本微整容图片数据集上进行经多次对照实

验，对比 3 种网络模型模型训练的结果。试验结果表明在此数据集上，HSANet (3 × 3)的各项指标较高，如表 1 是在微整容数据集上对比三种网络性能。

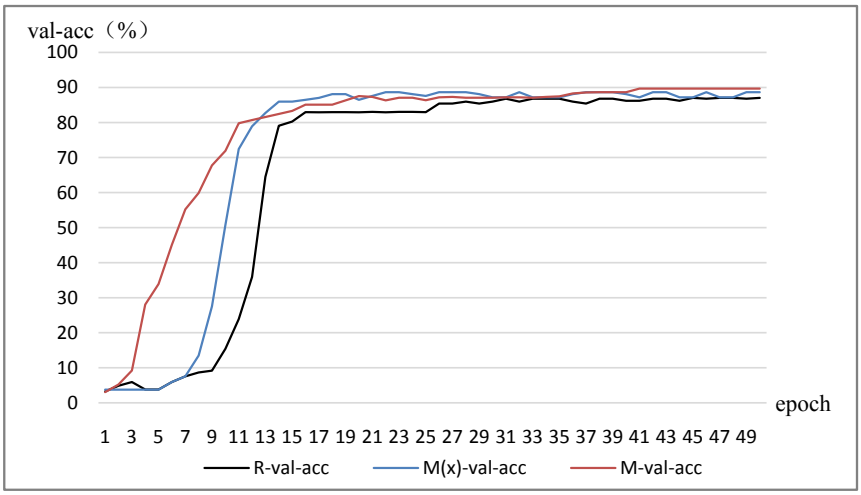

Figure 8. Verification set correct recognition rate comparison graph
图 8. 验证集正确识别率对比图

Table 1. Three kinds of network performance
表 1. 三种网络的性能

网络	f1-score	Macor avg	weight avg
ResNet50	0.50	0.50	0.50
HSANet (x)	0.53	0.52	0.54
HSANet (3 × 3)	0.60	0.58	0.59

为了说明该实验为何在 Identity 连接上用 3 × 3 卷积，先用改进得到得混合型自注意力块 HSANet (x)做对比，在该实验数据集上 HSANet (3 × 3)网络性能优于 HSANet (x)，选用现有的 ResNet50 和改进后的 HSANet (x)进行对照，在表 1 中 f1-score 为 f1 分数是衡量二分类模型精确度的一种指标[16]，即模型精确率和召回率的一种加权平均，macor avg 为宏平均是所有类的平均精准；weight avg 为加权平均对宏平均的一种改进，考虑每个类别样本数量在总样本中占比精准加权，评估模型性能的值越高表示网络模型性能越好[17]，3 种网络模型里 HSANet (3 × 3)的 3 项指标都明显高于其他 2 种网络模型。

可以发现，在 ResNet50 的 Bottleneck 中加入 Self-Attention 的时候，表现出更好的效果。2 组对照实验中，HSANet (3 × 3)内存及参数量相比 ResNet50 和 HSANet (x)多。如表 2 是三种网络模型在微整容数据集上的基准方法。

Table 2. Benchmark methods of the three networks
表 2. 三种网络的基准方法

网络	Forward/back pass size (MB)	Params size (MB)
ResNet50	286.55	89.88
HSANet(x)	300.76	66.64
HSANet(3×3)	302.35	97.54

4. 总结

针对微整容导致影响身份识别的问题，本文自制了微整容小样本图片数据集，本文创新点如下两点：

1) 制作 26 类微整容人脸图片数据集。

2) 构建 HSANet (3 × 3)混合型自我注意力网络模型。

HSANet (3 × 3)是受 Bottleneck Transformer 启发，在 ResNet50 的 Bottleneck 改进的混合型自我注意力块。HSANet (3 × 3)模型在 Bottleneck 块中将卷积与 Self-Attention 结合，并将 Identity 连接上的 1 × 1 改为 3 × 3 的卷积，为了与此作比较保留了 HSANet (x) Identity 连接的默认值(即 1 × 1 卷积)，在 Identity 连接上设置 *conv*3×3 与 BN 可以将输出连接至线性网络中的 Bottleneck 块，实现了块内自我注意力输出的高联特征，与连接捕获的局部特征的结合，提高模型对局部和全局信息的捕获能力。

虽然本文提出的混合型自我注意力模型，对微整容小样本图片数据集具有较高的识别能力，也有较好的特征捕获和识别能力，同时为之付出了相应的内存。为此，接下来微整容人脸识别工作的主要任务是在稳定提高正确识别率的前提下，减少模型空间成本，进一步提升网络性能。

基金项目

国家自然科学基金项目(62141206)；新疆自然科学基金项目(2019D01C337)；伊犁师范大学'学实高层次人才岗位(YSXSJS22002)；伊犁师范大学博士科研启动项(2020YSBS005)。

参考文献

[1] 苏丽, 孙雨鑫, 苑守正. 基于深度学习的实例分割研究综述[J]. 智能系统学报, 2021, 17(1): 16-31.

[2] 赵宣栋, 陈曦. 卷积神经网络在表情识别上的研究综述[J]. 计算机时代, 2022(4): 1-4.

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

[4] 张文秀, 朱振才, 张永合, 等. 基于残差块和注意力机制的细胞图像分割方法[J]. 光学学报, 2020, 40(17): 70-77.

[5] 王鸣展, 冀俊忠, 贾奥哲, 张晓丹. 基于跨尺度特征融合自注意力的图像描述方法[J]. 计算机科学, 2022, 49(10): 191-197.

[6] Luong, T., Pham, H. and Manning, C.D. (2015) Effective Approaches to Attention-Based Neural Machine Translation. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, 17-21 September 2015, 1412-1421. https://doi.org/10.18653/v1/D15-1166

[7] 李翔, 张涛, 张哲, 等. Transformer 在计算机视觉领域的研究综述[J]. 计算机工程与应用, 2023, 59(1): 1-14. http://kns.cnki.net/kcms/detail/11.2127.TP.20221009.1217.003.html

[8] Zhang, H., Wu, C., Zhang, Z., et al. (2022) ResNeSt: Split-Attention Networks. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), New Orleans, 19-20 June 2022, 2735-2745. https://doi.org/10.1109/CVPRW56347.2022.00309

[9] Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A. and Chen, LC. (2020) Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, JM., Eds., *Computer Vision—ECCV* 2020. *Lecture Notes in Computer Science*, Vol. 12349, Springer, Cham, 108-126. https://doi.org/10.1007/978-3-030-58548-8_7

[10] Bello, I., Zoph, B., Vaswani, A. and Shlens, J. (2019) Attention Augmented Convolutional Networks. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 3285-3294. https://doi.org/10.1109/ICCV.2019.00338

[11] Srinivas, A., Lin, T.-Y., Parmar, N., et al. (2021) Bottleneck Transformers for Visual Recognition. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 16514-16524. https://doi.org/10.1109/CVPR46437.2021.01625

[12] 张膑, 张运杰, 白明明. 结合 Shortcut Connections 结构的卷积稀疏编码图像去噪算法[J]. 科学技术与工程, 2021, 21(26): 11253-11262.

[13] 刘成攀, 吴斌, 杨壮. 基于联合损失和恒等映射的动态人脸识别[J]. 传感器与微系统, 2021, 40(9): 153-156.

[14] 王凤鸽. 基于轻量化与多尺度注意力融合的图像分类[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2021. https://doi.org/10.27389/d.cnki.gxadu.2021.002338

[15] 李欣栩. 基于深度学习的评论文本情感分析[D]: [硕士学位论文]. 南京: 南京信息工程大学, 2021. https://doi.org/10.27248/d.cnki.gnjqc.2021.000497

[16] 刘晋, 邓洪敏, 徐泽林, 杨洋. 面向目标识别的轻量化混合卷积神经网络[J]. 计算机应用, 2021, 41(S2): 5-12.

[17] 赵朗月, 吴一全. 基于机器视觉的表面缺陷检测方法研究进展[J]. 仪器仪表学报, 2022, 43(1): 198-219.