引用格式：ZENG Junying, CHEN Yucong, LIN Xihua, et al. An Ultra-lightweight Real-time Segmentation Network of Finger Vein Textures[J]. Acta Photonica Sinica, 2022, 51(2):0210009
曾军英,陈宇聪,林惜华,等.一种超轻量级指静脉纹络实时分割网络[J].光子学报,2022,51(2):0210009

一种超轻量级指静脉纹络实时分割网络

曾军英[1],陈宇聪[1],林惜华[1],秦传波[1],王迎波[1],朱京明[1],田联房[2],翟懿奎[1],甘俊英[1]

(1 五邑大学 智能制造学部,广东 江门 529020)
(2 华南理工大学 自动化科学与工程学院,广州 510640)

摘　要:现有的指静脉分割网络大多需要消耗极大内存和计算资源,难以直接部署到嵌入式平台上,大部分模型轻量化方法存在参数减小导致分割性能急剧下降、算力受限和实时性等问题。针对上述问题,本文提出了一种超轻量级指静脉纹络实时分割网络—SGUnet。首先,使用沙漏状的深度可分离卷积极大地减少基础模型参数,并采用轻量级高效注意力模块实现无降维的局部跨通道交互,提升网络分割性能。其次,为了解决部分特征图存在冗余的问题,使用Cheap operation来替代部分"懈怠"的卷积核,得到相似的特征图。最后,采用特征信息交互的方法,打开分组卷积的组间通道,解决了分组特征组之间信息不流通的问题。与传统Unet分割网络相比,最终的SGUnet模型参数量约为传统Unet分割网络的1%,Mult-Adds约为0.5%。在两个公开的手指静脉数据集SDU-FV、MMCBNU-6000上验证网络性能,结果表明SGUnet网络在分割性能上不仅优于大型分割网络Unet、DU-Net、R2U-Net,而且超越了经典轻量级改进模型squeeze-Unet、Mobile-Unet、shuffle-Unet、Ghost-Unet。SGUnet网络Accuracy、Dice、AUC分别达到94.11%、0.538 4、0.935 4,并且在NVIDIA嵌入式平台上指静脉纹络提取的测试速度高达0.27秒/张。

关键词:手指静脉分割;轻量级网络;嵌入式平台;模型压缩;实时分割网络;图像分割;卷积神经网络
中图分类号:TP391　　**文献标识码**:A　　**doi**:10.3788/gzxb20225102.0210009

0　引言

在众多生物特征识别技术中,手指静脉识别因其非接触式采集、活体识别、不易伪造、成本较低等优点,吸引了大量科研工作者的关注。手指静脉纹络提取是指静脉识别技术的关键步骤,直接影响后续指静脉特征提取、匹配和识别的准确度。虽然通过经典语义分割网络FCN[1]、SegNet[2]、RefineNet[3]等也能实现较好的指静脉纹络提取,但这些方法需要占用大量的存储空间和计算资源,难以有效地应用在当下的嵌入式平台和移动终端上,因此,设计轻量化深度神经网络架构是解决该问题的关键。

近些年来,轻量级神经网络模型的设计吸引了学术界和工业界的广泛关注,并提出了一系列轻量级模型[4-10]。Xception[4]将空间卷积和通道卷积分开,将两个方向上的相关性分开。SqueezeNet[5]将卷积层分为扩展层和压缩层,在压缩层实现对模型通道数的压缩。MobileNet[6]采用了深度可分离卷积,在提高性能的同时优化了模型的复杂度。MobileNetV2[7]颠覆了正残差的思想,提出了反向残差块。ShuffleNet[8]提出通

基金项目:国家自然科学基金(No. 61771347),广东普通高校人工智能重点领域专项(No. 2019KZDZX1017),广东省数字信号与图像处理技术重点实验室开放基金(Nos. 2019GDDSIPL-03,2020GDDSIPL-03),广东普通高校重点领域专项(No. 2020ZDZX3031),广东省基础与应用基础研究基金(Nos. 2021A1515011576, 2019A1515010716),2021 年度江门市基础与理论科学研究类科技计划项目(江科〔2021〕87号)
第一作者:曾军英(1977—),男,副教授,博士,主要研究方向为图像处理、深度学习理论与应用。 Email: zengjunying@126.com
通讯作者:秦传波(1982—),男,讲师,博士,主要研究方向为医学影像处理、生物特征识别。 Email: tenround@163.com
收稿日期:2021-08-17;**录用日期**:2021-10-13

道混洗的方法，解决了分组卷积中各组信息的交流问题。ShuffleNetV2[9]从模型实时性的角度思考，提出了网络运行速度更快的4个基本原则。GhostNet[10]首次从特征冗余的角度思考，提出ghost模块对模型进行压缩。此外，网络轻量化的方法还有神经网络模型的压缩方法，如知识蒸馏[11]、剪枝[12]、量化[13]、低秩分解[14]等，这些方法通过不同的角度对模型进行压缩，取得了丰硕的成果。基于神经网络架构搜索（Neural Architecture Search，NAS）的自动化神经网络架构设计，MobileNetV3[15]提出采用NAS方法搜索更高效的神经网络，MnasNet[16]和NasNet[17]通过强化学习方法学习神经网络架构搜索策略，实现便携式设备上轻量化神经网络的自动化构建。

构建轻量级网络模型实现参数量减小的同时确保网络性能基本不变或略微下降是现在所面临的重要难题。注意力模块是提升网络性能的有效方法，最早提出注意力机制的SE[18]模块利用全连接层将通道特征赋予权重以加强网络性能，CBAM[19]在通道和空间维度上实现不同维度注意力分离策略，同时关注空间和通道上的特征。后续研究工作如GENet [20]和TA[21]，通过采用不同的空间注意力机制或设计高级注意力块来深化这一方法的应用。A2Net[22]和CCNet[23]利用非本地机制来限制捕获不同类型的空间信息。CA[24]模块将频道注意力分解为两个并行的一维特征编码，将空间坐标信息整合到注意力中。但是大部分注意力模块在提升网络的性能同时引进大量参数，不利于网络轻量化，因此探索一种可以提升网络性能并使其引进参数可忽略不计的注意力模块是十分有意义的。

指静脉分割在嵌入式平台实现问题近似于一个多目标优化问题，要同时考虑到分割性能、网络参数大小与运行时间。为解决上述问题，本文提出了一种超轻量级指静脉纹络分割网络来实现指静脉纹络的提取。第一步，使用沙漏状的深度可分离卷积对网络进行初步轻量化，并引入轻量级高效通道注意力（Effificient Channel Attention，ECA）模块提高网络性能，并且该注意力模块所引进参数可以忽略不计。第二步，在网络中采用“低廉的操作”（Cheap operation）[10]，将部分特征图通过简单映射得到，使网络进一步压缩。第三步，采用分组卷积替代沙漏状的深度可分离卷积中的点卷积，并通过特征信息交互解决了分组卷积中各组信息之间无法流通的问题。最后，将网络部署到嵌入式平台上，并与其他模型进行了对比。

1 方法

1.1 网络结构

SGUnet网络是针对嵌入式平台实现实时指静脉纹络提取，即指静脉分割问题，需要综合考虑分割性能、网络参数大小和运行时间。因此在网络中采用多路分支减小模型的参数量或添加各种提高精度的模块，均会导致网络在嵌入式平台的运算速度大大降低。考虑到这些问题，本文将RONNEBERGER等[25]提出的Unet编码-解码的网络结构作为基础，提出一种超轻量级指静脉分割网络-SGUnet，如图1所示。具体方法为：1）利用沙漏状的深度可分离卷积构建一个新型的轻量分割网络SGUnetV1，在网络中保留了Unet编码-解码结构特点的同时，使模型初步轻量化。并且加入ECA模块实现无降维的局部跨通道交互，提升网络的分割性能；2）考虑到部分通道卷积“懈怠”的现象，即所提取的特征图存在冗余，针对这一问题利用Cheap operation生成部分特征图，进一步将模型进行压缩得到SGUnetV2，Cheap operation具体结构如1（c）；3）SGUnetV3作为最终的模型，采用特征信息交互的方式，打破了卷积层的通道相关性和空间相关性是可以退耦合的设想，将各组特征均匀地随机排列重组，解决了分组卷积组间信息的交流问题。

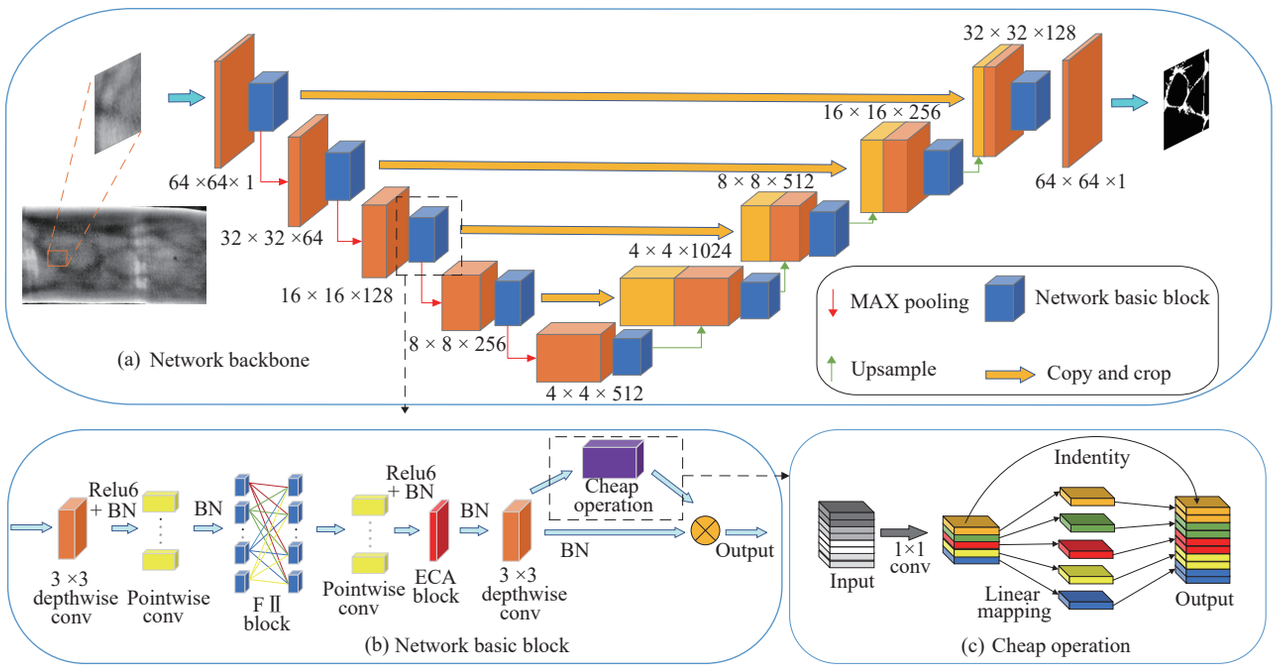


图 1　SGUnetV3 算法整体结构

Fig.1　The overall structure of the SGUnetV3

1.2　模型构建方法

1.2.1　沙漏状的深度可分离卷积和 ECA 轻量级注意力模块

第一步使用沙漏状的深度可分离卷积对网络进行初步轻量化，并加入注意力 ECA 模块提高模型性能。网络的基础模块如图 2(b)所示，由此基础块构建的模型称为 SGUnetV1。

沙漏状的深度可分离卷积：由 ZHOU Daquan[26] 等提出的 sandglass block 可以解决 Inverted residual

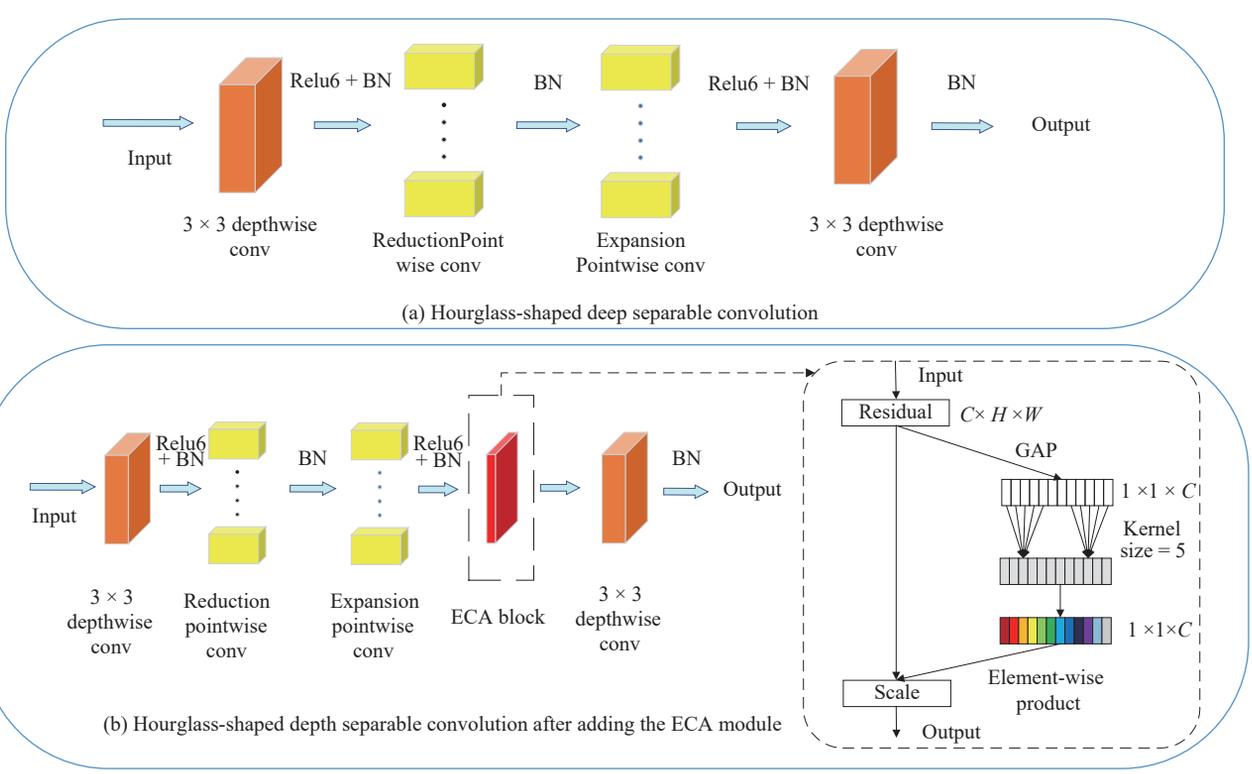


图 2　SGUnetV1 基础块结构图

Fig.2　SGUnetV1 basic block structure diagram

block[7]存在的特征信息不足的问题。本文采用sandglass block作为模型基础块，替代普通卷积提取特征，如图2(a)所示。假设输入图片为$\boldsymbol{X}_i \in R^{H\times W\times C_i}$，其中$H$，$W$，$C$分别为输入图片的长，宽和通道数。首先，利用3×3逐层卷积(depthwise conv)对输入特征图进行深度方向上的特征提取，此时并没有对输入$\boldsymbol{X}_i$进行维度上的压缩，所提取的空间特征具有更强的表达性。然后，将经过第一层逐层卷积后的特征图$\boldsymbol{X}' \in R^{H'\times W'\times C_i}$，作为沙漏状的点卷积层(1×1 pointwise conv)的输入，沙漏状的点卷积层由两个点卷积构成，其设计遵循正残差的结构，第一个点卷积先将特征维度按比例$\boldsymbol{r}$压缩，此时的输出为$\boldsymbol{X}'' \in R^{H'\times W'\times\left(\frac{C_i}{r}\right)}$，第二个点卷积将维度恢复至输入的特征维度$\boldsymbol{C}_i$，在提取到充足的通道特征信息的前提下，减小点卷积所带来的巨大开销。经过前面逐层卷积和点卷积之后，输入图片在深度和通道两个方向提取到丰富的特征，但在点卷积层中为了削减点卷积的开销，遵循了正残差的先压缩再扩展的思想，这会导致部分空间信息在缩小的点卷积处丢失，为了弥补这些特征的丢失，在最后再加入一层3×3逐层卷积层来补充空间的特征信息，弥补所丢失的部分空间信息。为了验证所使用的方法，将初步改进的模型SGUnet与Unet、Inverted residual block融入骨干Unet网络进行了对比实验，实现结果如表1所示，结果表明沙漏状的深度可分离卷积相比于普通卷积，逆残差块更轻量，效果更好。

由于考虑到在利用沙漏状的深度可分离卷积初步轻量化模型时，模型的分割性能可能会随着模型参数量的减少而下降，以及部分注意力模块存在降维操作破坏通道间与权重的直接对应关系，因此在沙漏状的深度可分离卷积的第二次点卷积(Expansion pointwise conv)之后和最后的3×3逐层卷积之间中加入了ECA注意力模块[28]，结构如图2(b)。

ECA轻量级注意力模块：SE注意力模块[18](Squeeze-and-Excitation)可以提升网络性能这一结论已被证实，但大部分注意力模块的加入在提升性能的同时也为网络增加大量运算负担。首先回顾SE注意力模块，令一个卷积块的输出为$\varphi_s \in R^{H\times W\times C}$，其中$W$，$H$和$C$为宽度、高度和通道尺寸，SE块中的信道权重计算为

$$w=\sigma\left(f_{\{\boldsymbol{W}_1,\boldsymbol{W}_2\}}\left(g(x)\right)\right) \tag{1}$$

式中，$g(x)=\frac{1}{WH}\sum_{i=1,j=1}^{W,H}x_{ij}$为通道全局平均池化，$\sigma$为sigmoid函数。为了避免较高的模型复杂度，将$\boldsymbol{W}_1$和$\boldsymbol{W}_2$的大小分别设置为$C\times\left(\frac{C}{r}\right)$和$\left(\frac{C}{r}\right)\times C$。在式(1)中$W_1$和$W_2$降维操作可以减少模型复杂性，但是它破坏了通道间和它的权重之间的直接对应关系。

综合考虑了性能提升和运算量增加两方面的取舍，以及注意力模块降维会破坏通道间与权重的直接对应关系，在SGUnet的网络结构中加入了ECA注意力模块[27]。ECA模块先将输入特征图通过全局平均池化(GAP)可用式(2)表示，将输入特征$\boldsymbol{\varphi}_s \in R^{H\times W\times C}$转化为$\boldsymbol{\varphi}_o \in R^{1\times 1\times C}$为

$$g(\boldsymbol{\varphi})=\frac{1}{WH}\sum_{i=1,j=1}^{W,H}\boldsymbol{\varphi}_{ij} \tag{2}$$

$$\boldsymbol{\varphi}_o=g(\boldsymbol{\varphi}_s) \tag{3}$$

然后经过计算量极低的一维卷积生成信道权重$\{\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{W}_3,\cdots,\boldsymbol{W}_c\}$

$$\boldsymbol{W}_j=\varepsilon\left(\sum_{i=1}^{k}\gamma^i y_j^i\right) \tag{4}$$

式中，$y_j^i \in \theta_i^k$，θ_i^k表示$\boldsymbol{\varphi}_o$中k个相邻通道集合，ε为激活函数sigmoid，其中γ为一维卷积操作。最后通过信道权重与输入特征图的乘积得到输出特征图，在提升网络性能的同时使运算增加量可忽略不计。

1.2.2 从特征图冗余的方向进行模型压缩

在SGUnetV1的基础上，将Cheap operation置于沙漏状的深度可分离之后，从特征图冗余的角度进一步压缩了模型参数量。此时的网络基础块如图3所示，第二步改进后得到的网络称为SGUnetV2。

Cheap operation：绝大多数卷积神经网络没有考虑到特征图可能存在一定的冗余，这些冗余的特征图存在极大的相似性，它们可以通过一些简单的变化从相似的特征图中得到。HAN K等[10]提出的GhostNet正是从特征图冗余的角度思考，对模型进行压缩。在采用深度可分离卷积的网络架构中，大量的1×1卷积还

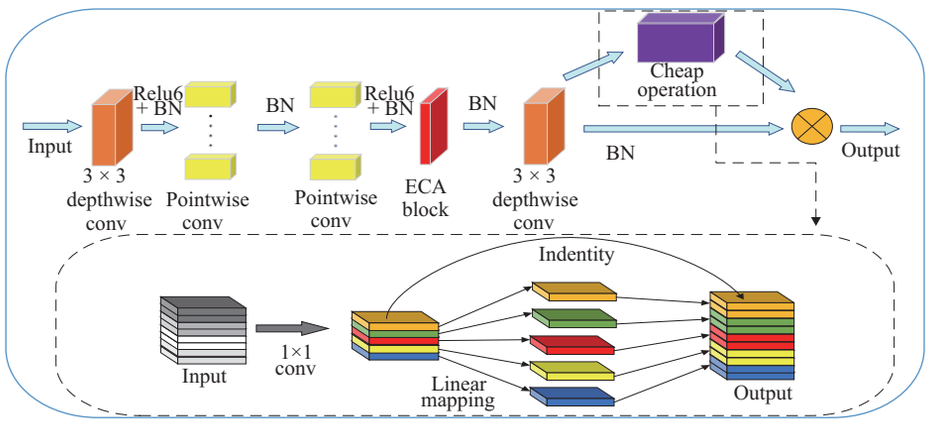


图3　加入Cheap operation后的基础模块结构

Fig.3　The basic module structure after joining the Cheap operation

会占用庞大的MAC(Memory Access Cost)和参数量。Ghostnet中所提出的Ghost模块，从特征图冗余的角度，将两张十分相似特征图的其中一张看作另一张特征图的“影子”，这类特征图可以通过Cheap operation所得，从而减少大量1×1卷积。

Cheap operation是针对1×1卷积的压缩方式，此处针对1×1卷积运算削减量进行分析。对于输入特征图$\boldsymbol{X}_{\mathrm{i}}$可将其表示为$H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}$，所需的输出特征图$\boldsymbol{X}_{\mathrm{o}}$可表示为$H_{\mathrm{o}}\times W_{\mathrm{o}}\times C_{\mathrm{o}}$，由于1×1卷积仅改变特征图的维度，因此不采用Cheap operation需要C_{o}个$1\times 1\times C_{\mathrm{i}}$卷积核，此时得到目标输出特征图所需要的计算量为$C_{\mathrm{o}}\times H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}$。如果采用Cheap operation对1×1卷积按比例K进行压缩，1×1的计算量则有$\dfrac{C_{\mathrm{o}}\times H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}}{K}$，需要线性变换的计算量为$\dfrac{(K-1)\times C_{\mathrm{o}}}{K}$。此时Cheap operation的运算量为$\dfrac{(K-1)\times C_{\mathrm{o}}}{K}+\dfrac{C_{\mathrm{o}}\times H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}}{K}$。实际压缩的参数之比为

$$\frac{C_{\mathrm{o}}\times H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}}{\dfrac{(K-1)\times C_{\mathrm{o}}}{K}+\dfrac{C_{\mathrm{o}}\times H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}}{K}}=\frac{K\times H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}}{(K-1)\times H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}} \tag{5}$$

由于在浅层的网络中输入特征图的大小$H_{\mathrm{i}}, W_{\mathrm{i}}\gg K$，在深层的网络中输入通道数往往成百上千，即$C_{\mathrm{i}}\gg K$，此时可将式(5)近似于

$$\frac{K\times H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}}{(K-1)+H_{\mathrm{i}}\times W_{\mathrm{i}}\times C_{\mathrm{i}}}\approx K \tag{6}$$

将Cheap operation用作针对点卷积参数量减小的方法，每层网络都采用了该方法得到部分特征图，进一步减小网络大小。

1.2.3　分组卷积和特征信息交互

第三步，在利用分组卷积减轻1×1卷积负担的同时，采用特征信息交互的方式来弥补分组卷积缺乏组间信息交互的缺点，具体操作如图4(b)所示，图4(a)所展示的为添加Channel shuffle[8]后的网络基础块，将由图4(a)构建的网络称其为SGUnetV3。

特征信息交互：即使经过第二步的改进，网络运算量和参数量依然大量堆积在1×1卷积处，ZHANG X等[8]采用分组卷积来减轻1×1卷积计算量的同时提出Channel shuffle解决了分组卷积之间每组特征信息无法交互的问题。在SGUnetV2的基础上，本文也采用了相同的方法，将分组卷积替代沙漏状的深度可分离卷积中的1×1卷积，非常可观地减小了1×1卷积的花费，并且采用特征信息交互的方法解决分组卷积中每组特征间缺乏信息交互的问题，在进一步压缩模型的同时，保证模型的性能。

Channel shuffle设计思路具体如下，假设使用S组分组卷积来替换昂贵的1×1卷积，为了解决组间信息传递的问题，将每组的分组卷积提取到的通道特征随机地分为S'组($S'=S$)，记为$\eta_j=\dfrac{1}{S'}, j\in[1,S]$，然后通

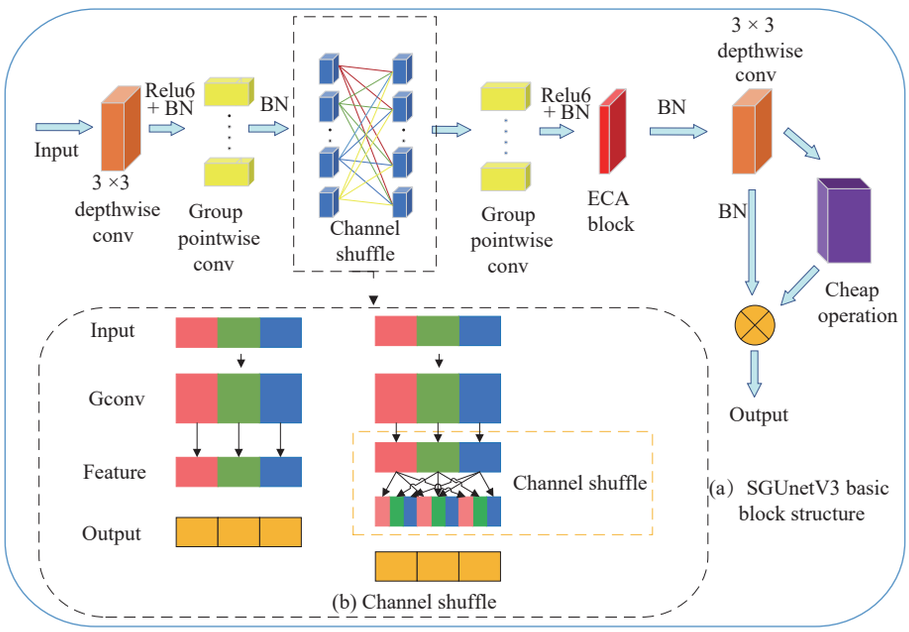


图 4 SGUnetV3 网络基础块结构

Fig.4 The basic block structure of SGUnetV3 network

过特征信息交互的方法，将随机选取每组划分后的其中一组特征 η_j，其他各组的操作也相同，最后将提取到的各组特征相叠加后进行特征转置，使从 S 组中得到小组信息充分融合。具体实施方法如图 4(b)所示。图 4(b)左侧为传统的分组卷积，每种颜色代表一组分组卷积，右侧为采用 Channel shuffle 后的分组卷积，解决了各组之间特征信息无法交互的问题。

2 实验

2.1 数据预处理

本文算法所采用的公开指静脉数据库为山东大学 SDU-FV 和韩国全北国立大学 MMCBNU_6000。将 SGUnet 与不同的分割网络进行对比，并且在嵌入式平台上进行对比实验。SDU-FV 数据集共有 106 个受试者，分别收集了每个人左右手的食指、中指和无名指三个手指的指静脉图像，每个手指采集 6 张图片。因此，该库中共有 636 类(106 人×6 个手指)3 816(106 人×6 个手指×6 样本)张图像。MMCBNU_6000 数据集共有 100 个受试者，分别采集了每个人左右手的食指、中指、无名指三个手指的指静脉图像，每个手指采集 10 张图片，且该数据集给出了提取好的 ROI 区域。因此，该库中共有 600 类(100 人× 3×2 个手指) 6 000 (600 人×6 个手指×10 样本指)张图像。

在训练的过程中随机选取数据集的五分之四作为训练集，剩余五分之一作为测试集。在训练和测试中对原图采用分块策略，每张图像分为 2 000 个块(patch)，在宽和高均为五个步幅的情况下，对每张图像提取多个连续的重叠块，通过对覆盖像素的所有预测块的概率进行平均估计，从而获得该像素是静脉血管的概率。为了保证不超过硬件平台的内存限制及实时性，在指标与时间中权衡选取步幅为 5 的 patch 最为合适，在网络输出 patch 结果后按照分 patch 中的顺序，采用交叠滑窗策略保留中心区域结果，舍弃预测不准的图像边缘，重新拼接成一张完整的原图。

2.2 实验环境配置与嵌入式平台

为了展示本文方法的高效性和普适性，本文分别在 PC 端和嵌入式平台进行对比实验。PC 端对比实验的运行环境为 Ubuntu18.04 系统、英特尔 i9-10900k@3.7GHz CPU(10 核 20 线程)、内存 32GB、显卡 Nvidia GeForce RTX 3090 TURBO(24GB/技嘉)、CUDA11.2、Pytorch1.8.1、Python3.6.5。使用 Adam 优化器进行梯度下降，学习率为 0.001，batch size 大小为 512。

此外，在 NVIDIA 全系列的嵌入式平台 JETSON NANO、JETSON TX2、JETSON XAVIAR NX、

JETSON AGX XAVIAR上验证本文方法的普适性（嵌入式平台算力顺序：NANO＜TX2＜NX＜AGX）。为了让在不同嵌入式平台上的数据更具有对比性，将所有嵌入式平台环境配置设为一致Jetpack4.4、pytorch1.8.0、python3.6.9。从实验数据中可以证明该方法相较于以往的方法更高效、更轻量。

2.3 网络性能评价指标

在对比实验中，本文采用Dice，AUC，Accuracy，Specificity，Precision 5个分割指标作为评估网络性能优劣的依据，其中以Dice、AUC和Accuracy作为最主要的评估指标，在本小节中将简单介绍所采用的评估值标。

Dice是指使用频率最高的分割指标，分割结果和标注（ground truth）两个相交面积占总面积的比值，完美分割为1。AUC（Area Under roc Curve）的值为处于ROC（Receiver Operating Characteristic）曲线下方面积的大小。指标越接近1代表分割出来的性能越好。Accuracy是指对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。Specificity是指在实际为正确的样本中正确判断的概率。Precision是指在分类为正确的样本中正确判断的概率。

2.4 网络实验对比

在网络改进的第一步中，采用沙漏状的深度可分离卷积对基础网络进行初步的轻量化，得到SGUnet，并且为了验证该网络的高效性，将SGUnet与经典的Unet，用逆残差改进后的MobileV2＋Unet进行了对比，结果表明SGUnet的模型表现优异。与Unet相比，SGUnet的参数量约为Unet的3.8%，Mults-Adds约为Unet的2%，Dice，AUC和Precision三个评价指标显著提升，提升量分别为5.84%，5.48%，3.48%，Accuracy也提升了0.73%。与MobileV2＋Unet相比，SGUnet的模型参数量和Mults-Adds分别只占其1/10和1/4，在分割性能上SGUnet大大超越MobileV2＋Unet，具体实验数据如表1所示。

表1 改进网络SGUnet与基础Unet、MobileV2+Unet性能对比表

Table 1 Improved network SGUnet and basic Unet, MobileV2+Unet performance comparison table

Network	Params	Mult-Adds	Dice	AUC	Accuracy	Specificity	Precision
Unet	13.39M	1.928G	0.444 6	0.843 4	91.16%	96.47%	53.79%
MobileV2＋Unet	5.289M	171.226M	0.502 5	0.855 4	91.31%	95.34%	53.68%
SGUnet	516.014k	39.494M	0.503 0	0.898 2	91.89%	95.95%	57.27%

在SGUnet的基础上加入ECA模块提升了模型性能，将所采用的ECA模块与具有代表性的SE模块，最新的CA模块[24]分别用于SGUnet模型上进行对比实验，以此来验证所加入ECA模块更适合在SGUnet网络中使用。此外，本文还探究了各个模块中设置不同的参数对实验结果的影响。实验结果如表2所示。（CA模块后的数字为模块压缩尺度设置，ECA模块后的数字为1D kernel size的尺寸）。

由于考虑到ECA模块放置不同也会导致网络性能有所影响，为此本文设计了a，b，c三种方案，见图5，并进行了大量的对比实验，最终发现采用a设计方案的网络性能更好，具体设计方案性能对比见表3。

表2 在SGUnet的基础上加入ECA模块与经典SE模块、CA注意力模块的性能对比表

Table 2 Performance comparison table of ECA module, classic SE module and CA attention module added on the basis of SGUnet

Network	Dice	AUC	Accuracy	Specificity	Precision
SGUnet	0.503 0	0.898 2	91.89%	95.95%	57.27%
SGUnet＋SE	0.496 1	0.892 6	91.95%	96.32%	58.27%
SGUnet＋CA-32	0.496 8	0.886 7	91.89%	96.25%	57.79%
SGUnet＋CA-16	0.500 8	0.891 7	91.93%	96.18%	57.87%
SGUnet＋CA-8	0.501 5	0.886 9	92.00%	96.30%	58.55%
SGUnet＋ECA-3	0.497 8	0.891 3	91.80%	96.11%	58.37%
SGUnet＋ECA-5	0.503 8	0.898 2	92.10%	96.34%	59.04%

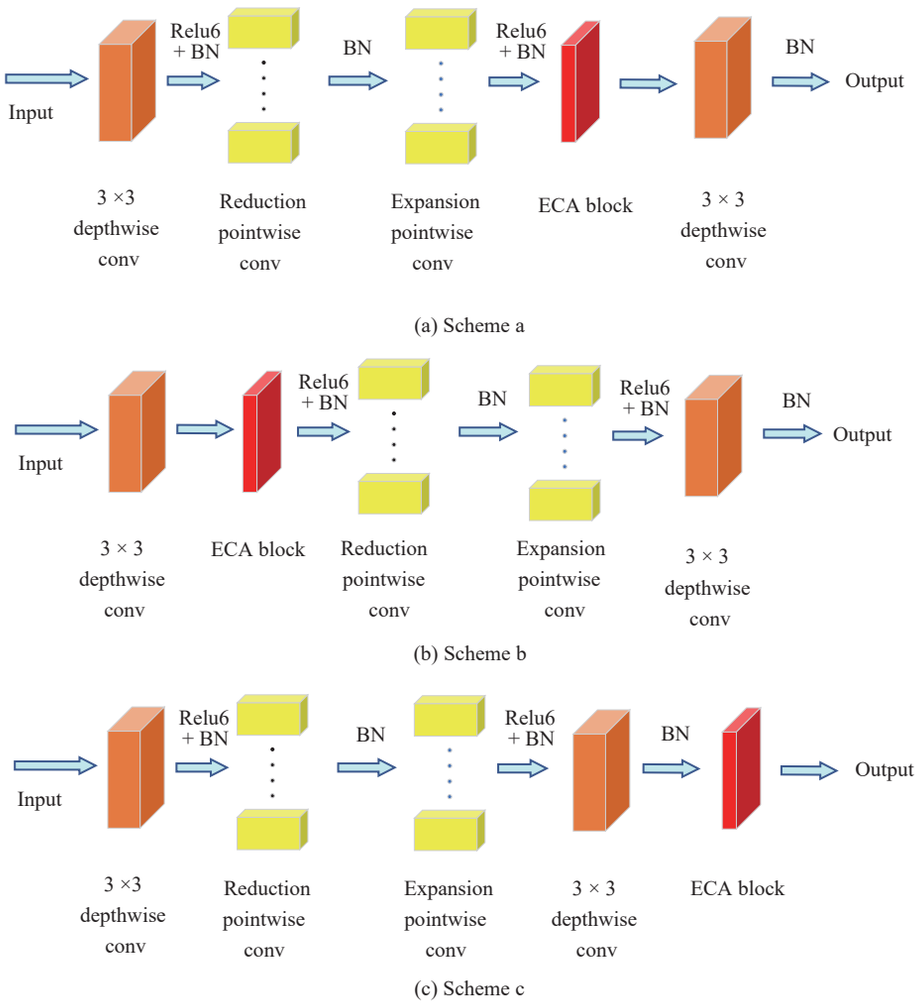


图 5 ECA模块放置的三种方案

Fig.5 Three options for ECA module placement

表 3 a,b,c三种不同方案的网络性能比较

Table 3 Comparison of network performance of three different schemes a, b, and c

Network	Dice	AUC	Accuracy	Specificity	Precision
a	0.501 6	0.898 2	92.10%	96.34%	59.04%
b	0.501 2	0.896 4	92.03%	96.18%	58.54%
c	0.500 4	0.894 5	91.97%	96.26%	58.32%

在网络的第二步改进中,使用Cheap operation替代部分“懈怠”卷积操作,进一步减轻网络的大小以及运算量,此时将所得到的SGUnetV2与Unet,用深度可分离卷积改进的MobileV1+Unet,使用逆残差改进的MobileV2+Unet在模型大小上进行对比见表4,在SGUnetV1的基础上模型参数减小100k,Mult-Adds减小13M,在模型压缩上我们保持十分明显的优势。

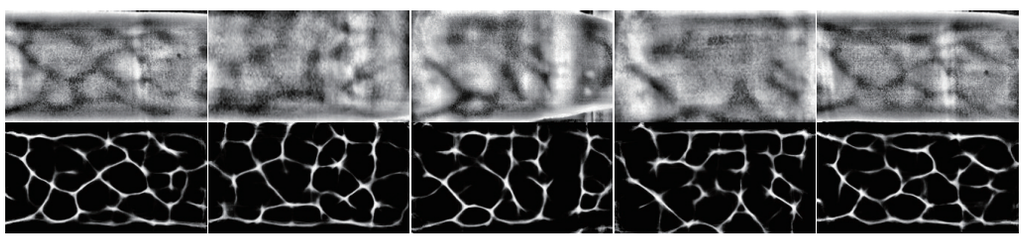
表 4 SGUnetV2与其他网络的参数比较表

Table 4 Comparison of parameters between SGUnetV2 and other networks

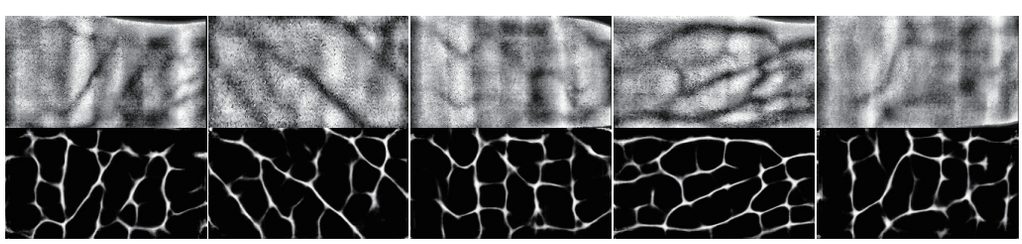
Network	Params	Mult-Adds	Dice	AUC	Accuracy	Specificity	Precision
Unet	13.39M	1.928G	0.444 6	0.843 4	91.16%	96.47%	53.79%
MobileV1+Unet	3.932M	481.35M	0.498 9	0.855 4	91.31%	95.34%	53.68%
MobileV2+Unet	5.289M	171.226M	0.502 5	0.884 6	91.54%	95.87%	56.68%
SGUnetV1	516.054k	39.504M	0.503 8	0.898 2	92.10%	96.34%	59.04%
SGUnetV2	416.752k	26.575M	0.499 2	0.898 9	91.73%	96.12%	56.14%

2.5 模型可视化

为了更直观地展示网络模型的分割效果,本节展示了在两个指静脉数据集上的分割效果图,其中每个数据库各选取五张图进行结果展示,如图6,并且在图7中展示了在将每张原图分成多张patch后,单张patch在网络中的可视化效果,进一步验证网络模型对指静脉的细节分割效果较好。



(a) SDU-FV

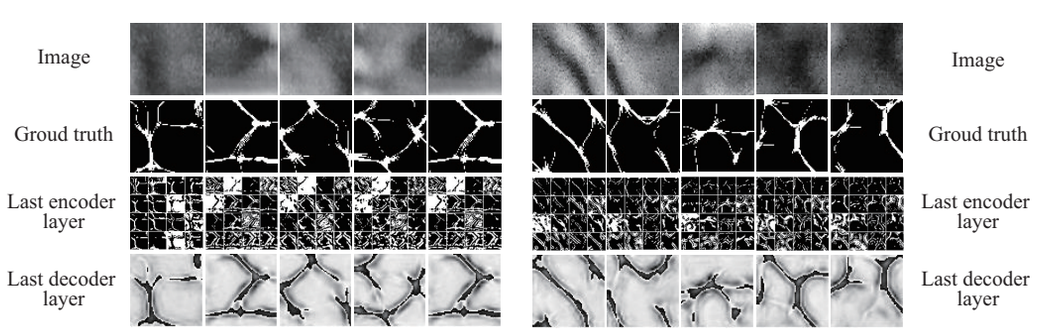


(b) MMCBNU-6000

图6 指静脉分割效果图

Fig. 6 The effect of finger vein segmentation

图7为单张patch在SGUnetV3网络中的可视化效果。其中Imgae为原图分成的patch,Groud truth为单张patch所对应的标签块,Last encoder layer展示了最后一层编码器中所提取到patch中的指静脉的边界、轮廓等深层特征,Last decoder layer为网络最后一层解码器中所提取到指静脉纹络。对比网络最后一层解码器与其所对应的标签块,可以明显地看出所提取到的指静脉与标签几乎无异。



(a) SDU-FV (b) MMCBNU_6000

图7 指静脉每个patch分割可视化

Fig.7 Segmentation visualization of each patch of finger vein

2.6 与大型网络在PC端的对比实验

在PC端将SGUnet系列网络与同基础的网络U-net、变更的网络R2U-net以及DU-net在SDU-FV和MMCBNU_6000两个数据集上进行了对比实验。本文采用常用的图像分割评判指标Dice系数、准确性(Accuracy)、特异性(Specificity)、精确率(precision)、AUC(Area Under roc Curve,AUC为ROC curve下方与x轴围成的面积大小)来判断网络性能的优劣。实验结果如表5和表6。

从实验数据中可以观察得到所提出的网络SGUnetV3与大型网络R2UNet,DUnet,以及Unet在模型大小上的比值分别为0.2%,0.5%,1%。SGUnetV3在参数量,运算时间和计算量上极具优势,并且SGUnetV3在性能方面相比Unet有显著提升。

在SDU-FV数据集上,SGUnetV1,V2,V3分割性能得到全面提升,分割系数Dice和AUC提升显著,证

表5 SGUnet系列网络与大型网络在SDU-FV数据集上的实验结果

Table 5 Experimental results of SGUnet series network and large-scale network on SDU-FV data set

Network	Params	Mult-Adds	Dice	AUC	Accuracy	Specificity	Precision
R2UNet	48.92M	------	------	0.902 9	91.87%	98.21%	62.18%
DUNet	26.73M	------	------	0.913 3	91.99%	97.26%	64.20%
Unet	13.39M	1.928G	0.444 6	0.843 4	91.17%	96.48%	53.79%
SGUnetV1	516.054k	39.504M	0.503 8	0.898 2	92.10%	96.34%	59.04%
SGUnetV2	416.752k	26.575M	0.499 2	0.898 9	91.73%	96.12%	56.14%
SGUnetV3	145.25k	10.453M	0.497 3	0.899 2	91.60%	95.81%	55.34%

表6 SGUnet系列网络与大型网络在MMCBNU_6000数据集上的实验结果

Table 6 Experimental results of SGUnet series network and large-scale network on MMCBNU_6000 data set

Network	Params	Mult-Adds	Dice	AUC	Accuracy	Specificity	Precision
R2UNet	48.92M	------	------	0.905 8	92.94%	97.22%	54.68%
DUNet	26.73M	------	------	0.912 5	93.30%	97.89%	58.82%
Unet	13.39M	1.928G	0.437 2	0.847 4	91.03%	95.70%	49.49%
SGUnetV1	516.054k	39.504M	0.538 4	0.934 4	94.11%	96.79%	60.44%
SGUnetV2	416.752k	26.575M	0.527 9	0.935 4	93.75%	96.31%	57.55%
SGUnetV3	145.25k	10.453M	0.520 2	0.933 3	93.68%	96.36%	57.23%

明本文改进的方法是卓有成效的。SGUnetV3网络与大型经典网络R2UNet,DUnet相比,各项指标接近,由于模型在压缩过程中存在性能下降的情况,这是不可避免的,考虑到在网络大小和模型参数上的这一巨大差距以及在嵌入式平台上实现,可以认为SGUnetV3网络是高效的。

在MMCBNU_6000数据集上,SGUnet系列网络在性能方面超过R2UNet,DUnet。SGUnetV1的Dice系数达到了最高的0.538 4,每一步改进后网络的AUC指标都远超大型网络。SGUnetV3中AUC,Accuracy两个指标都有很大的提升,超过了大型的R2Unet,DUnet,Precision指标也仅稍稍低于DUnet,与R2Unet,DUnet在Specificity指标上相比整体相差较小,由于SGUnetV3模型大小与大型网络相差巨大,并且在部分指标上超越了大型网络,与基础的Unet相比,SGUnetV3网络大幅提升分割的性能,还将模型大小压缩至145k,这一结果也证明了网络的高效性。

2.7 与轻量化网络的对比实验

为了证明本文方法的实时性,本文在计算能力较弱的NVIDIA系列的嵌入式平台上部署SGUnet系列网络,并进行了测试实验。此外,本文还将一些经典的轻量化网络用在传统的Unet上与本文方法从模型大小、分割性能和运行时间等多个方面进行综合比较。实验结果如表7、表8所示。

表7 SGUnet系列网络与其它轻量级网络在SDU-FV数据集上的实验数据

Table 7 Experimental data of SGUnet series network and other lightweight networks on SDU-FV dataset

Network	Params	FLops	Mult-Adds	Dice	AUC	Accuracy	Specificity	Precision
Unet	13.39M	1.95G	1.928G	0.444 6	0.843 4	91.17%	96.48%	53.79%
Squeeze_Unet	2.893M	296.05M	287.61M	0.501 7	0.863 0	91.02%	94.72%	51.90%
Mobile_Unet	3.932M	498.13M	481.35M	0.502 5	0.855 4	91.31%	95.34%	53.68%
Ghost_Unet	6.783M	130.46M	128.97M	0.485 3	0.886 4	91.84%	96.75%	58.47%
Shuffle_Unet	516K	68.27M	57.97M	0.511 6	0.885 9	91.48%	95.28%	54.49%
SGUnetV1	516.054k	42.97M	39.504M	0.503 8	0.898 2	92.10%	96.34%	59.04%
SGUnetV2	416.752k	29.26M	26.575M	0.499 2	0.898 9	91.73%	96.12%	56.14%
SGUnetV3	145.25k	13.13M	10.453M	0.497 3	0.899 2	91.60%	95.81%	55.34%

2.7.1 模型大小

经过初步改进的SGUnetV1已经超越了经典的轻量级模型Squeeze_Unet,Mobile_Unet,Ghost_Unet,虽然在参数量上Shuffle_Unet与SGUnetV1相当,但是SGUnetV1在Flops、Mult-Add两项重要指标上分别比

表8　SGUnet系列网络与其它轻量级网络在MMCBNU_6000数据集上的实验数据

Table 8　Experimental data of SGUnet series network and other lightweight networks on MMCBNU_6000 dataset

Network	Params	FLops	Mult-Adds	Dice	AUC	Accuracy	Specificity	Precision
Unet	13.39M	1.95G	1.928G	0.474 1	0.883 4	92.42%	95.80%	49.24%
Squeeze_Unet	2.893M	296.05M	287.61M	0.510 5	0.921 6	93.08%	95.34%	52.60%
Mobile_Unet	3.932M	498.13M	481.35M	0.500 3	0.905 4	92.06%	94.52%	53.37%
Ghost_Unet	6.783M	130.46M	128.97M	0.511 0	0.924 3	93.01%	96.43%	58.38%
Shuffle_Unet	516K	68.27M	57.97M	0.479 2	0.906 6	91.80%	94.15%	46.33%
SGUnetV1	516.054k	42.97M	39.504M	0.538 4	0.934 4	94.11%	96.79%	60.44%
SGUnetV2	416.752k	29.26M	26.575M	0.527 9	0.935 4	93.75%	96.31%	57.55%
SGUnetV3	145.25k	13.13M	10.453M	0.520 2	0.933 3	93.68%	96.36%	57.23%

Shuffle_Unet减小了25.3MFlops、18.46M。由此可见SGUnetV1模型大小在第一步就超越了众多轻量级模型。然后，经过第二步利用Cheap operation代替SGUnetV1中部分懈怠的卷积核，得到了进一步的网络SGUnetV2，使模型参数降低至416.752k，Flops为29.26M，Mult-Add为26.575M。最后，通过利用分组卷积和特征信息交互，使网络再次压缩得到最终的SGUnetV3，其参数量只有145.25k，Flops仅为13.13M，Mult-Add只有10.453M。

2.7.2　分割性能

本文在两个公开的手指静脉数据集上验证了SGUnet系列网络高效性。SGUnetV1的性能超过了各种轻量级模型以及Unet，在SDU-FV数据集中，Accuracy指标达到了92.10%，AUC达到了0.898 2，Dice系数为0.503 8；SGUnetV2进一步压缩模型大小的同时，保证了网络的分割性能，其各项指标与SGUnetV1十分相近，AUC指标更是超越了SGUnetV1；最终的模型SGUnetV3只有145.25k参数，其AUC指标得到了最高的0.899 2，在其它指标上虽然略差于V1，V2，但也超过以往的轻量级网络和Unet，这一突破是巨大的。在MMCBNU_6000数据集上，SGUnetV1全面超越了所有的轻量级网络，相比于基础Unet，SGUnetV1的AUC，Dice，Accuracy这3个关键指标分别增涨了0.051，0.064，1.69%；SGUnetV2，V3也仅仅在Specificity，Precision指标上相比于Ghost_Unet稍弱，其它性能指标上也远超过了其他网络。

2.7.3　实际分割效果

图8、9分别为SGUnet，Unet和各种轻量级网络的分割结果图，在两个数据集中分别选取同一张图片对网络分割性能进行比较，图8中Unet分割效果并不理想，Squeeze_Unet，Mobile_Unet，Ghost_Unet和Shuffle_Unet的分割效果有所改进，但是部分极细的血管并没有很好地分割出来，以及分割得到的血管整体效果不够平滑。网络SGUnetV1不仅在性能指标上获得了出色的表现，实际分割得到的血管也更平滑，更符合人体血管的特点，从SGUnetV3的分割图中所标注的红框可以看出，部分极细的毛细血管也可以完美地分割。所提出的SGUnetV1、V2和V3在MMCBNU_6000数据集上也同样展现了强大的分割效果，可以清楚地从图9中观察到，其它轻量级网络在细节分割方面并没有优势，虽然它们都得到了比Unet更好的分割结果，但是在血管平滑程度和毛细血管的准确分割这两个方面还较为不足，SGUnet正好满足了这两个优势，图9红框所标注多个位置处的毛细血管被精准地分割，再次证明了SGUnet网络的高效性。

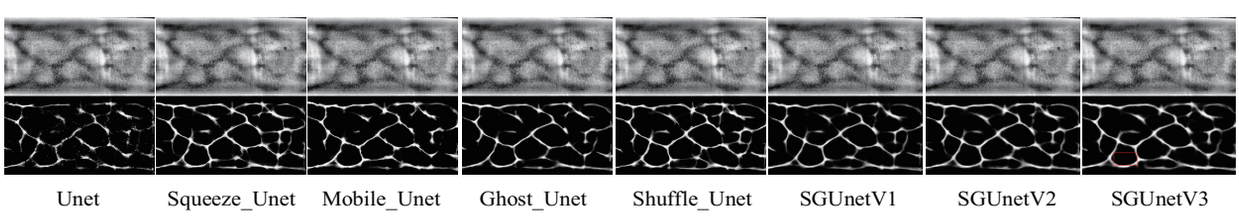


图8　SGUnet与各个轻量化网络在SDU-FV数据集上的实际分割效果图

Fig.8　The actual segmentation effect diagram of SGUnet and each lightweight network on the SDU-FV dataset

2.7.4　实时性实验

表9、10中分别记录了SGUnet系列网络与各种轻量化网络在NVIDIA全系列嵌入式平台上分割单张指静脉图片的运行时间，本文分别在SDU-FV和MMCBNU_6000两个数据集中选取20张图片进行测试。在SDU-FV数据集中，SGUnetV1仅在NVIDIA NANO上的运行时间略慢于Ghost_Unet，在其他嵌入式平台上运行速度完全超越了各种轻量化网络模型。虽然SGUnetV2，SGUnetV3与V1相比速度有所下降，这

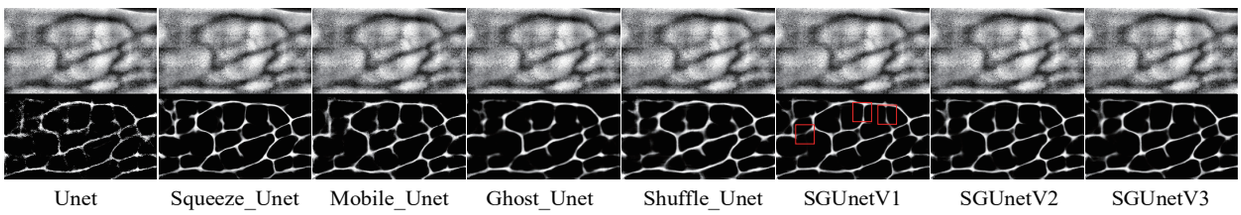


图 9 SGUnet与各个轻量化网络在MMCBNU_6000数据集上的实际分割效果图

Fig.9 The actual segmentation effect diagram of SGUnet and each lightweight network on the MMCBNU_6000 dataset

是由于网络轻量化引入了相比卷积更为复杂的操作，但V2，V3更加轻量，V3模型大小不到V1的1/3，我们认为这是可以接受的。在MMCBNU_6000数据集中，SGUnetV1在所有的嵌入式平台上的运行时间超越了所有的轻量化模型，证明SGUnet系列网络在MMCBNU_6000数据集上具有更好的效果。

表 9 SGUnet系列网络与其他轻量级网络在NVIDIA嵌入式平台上处理单张SDU-FV数据集图片的运行时间

Table 9 The running time of SGUnet series and other lightweight networks to process a single SDU-FV data set image on the NVIDIA embedded platform

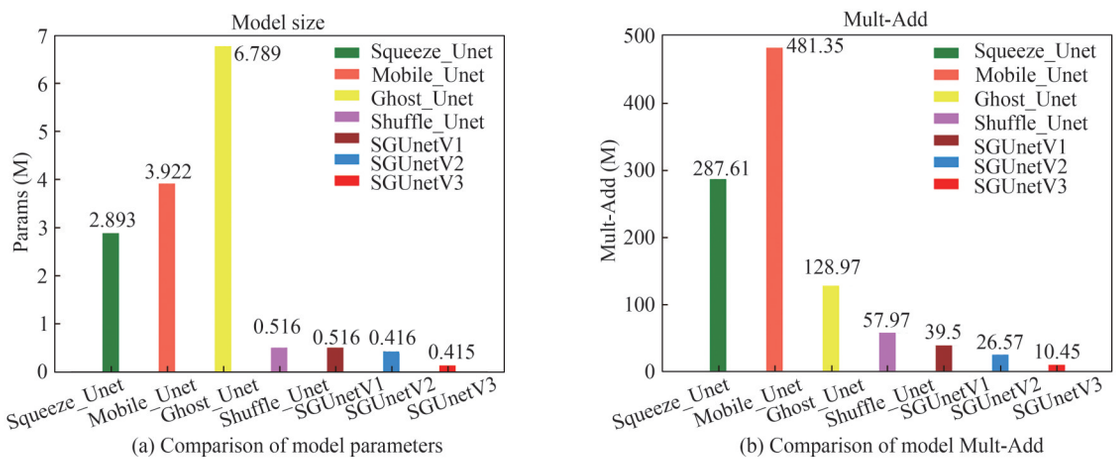
Network	Time/s			
	NANO	TX2	NX	AGX
Squeeze_Unet	5.146	0.686	0.663	0.388
Mobile_Unet	5.699	0.519	0.505	0.288
Ghost_Unet	2.719	0.684	0.722	0.358
Shuffle_Unet	3.208	0.766	0.654	0.426
SGUnetV1	2.735	0.455	0.386	0.283
SGUnetV2	2.789	0.523	0.427	0.302
SGUnetV3	2.827	0.569	0.458	0.306

表 10 SGUnet系列网络与其他轻量级网络在NVIDIA嵌入式平台上处理单张MMCBNU_6000数据集图片的运行时间

Table 10 The running time of SGUnet series and other lightweight networks to process a single MMCBNU_6000 data set image on the NVIDIA embedded platform

Network	Time/s			
	NANO	TX2	NX	AGX
Squeeze_Unet	5.241	0.679	0.699	0.357
Mobile_Unet	5.435	0.524	0.737	0.284
Ghost_Unet	2.791	0.612	0.734	0.346
Shuffle_Unet	3.221	0.766	0.652	0.418
SGUnetV1	2.779	0.457	0.405	0.270
SGUnetV2	2.873	0.525	0.418	0.290
SGUnetV3	2.928	0.570	0.467	0.291

由图10可以得到SGUnet系列模型从模型大小和性能指标两方面都胜于以往的经典轻量化方法。在图10(a)、(b)中可得，SGUnetV3模型远小于以往的轻量化模型，证明本文的轻量化方法效果更好；从



(a) Comparison of model parameters (b) Comparison of model Mult-Add

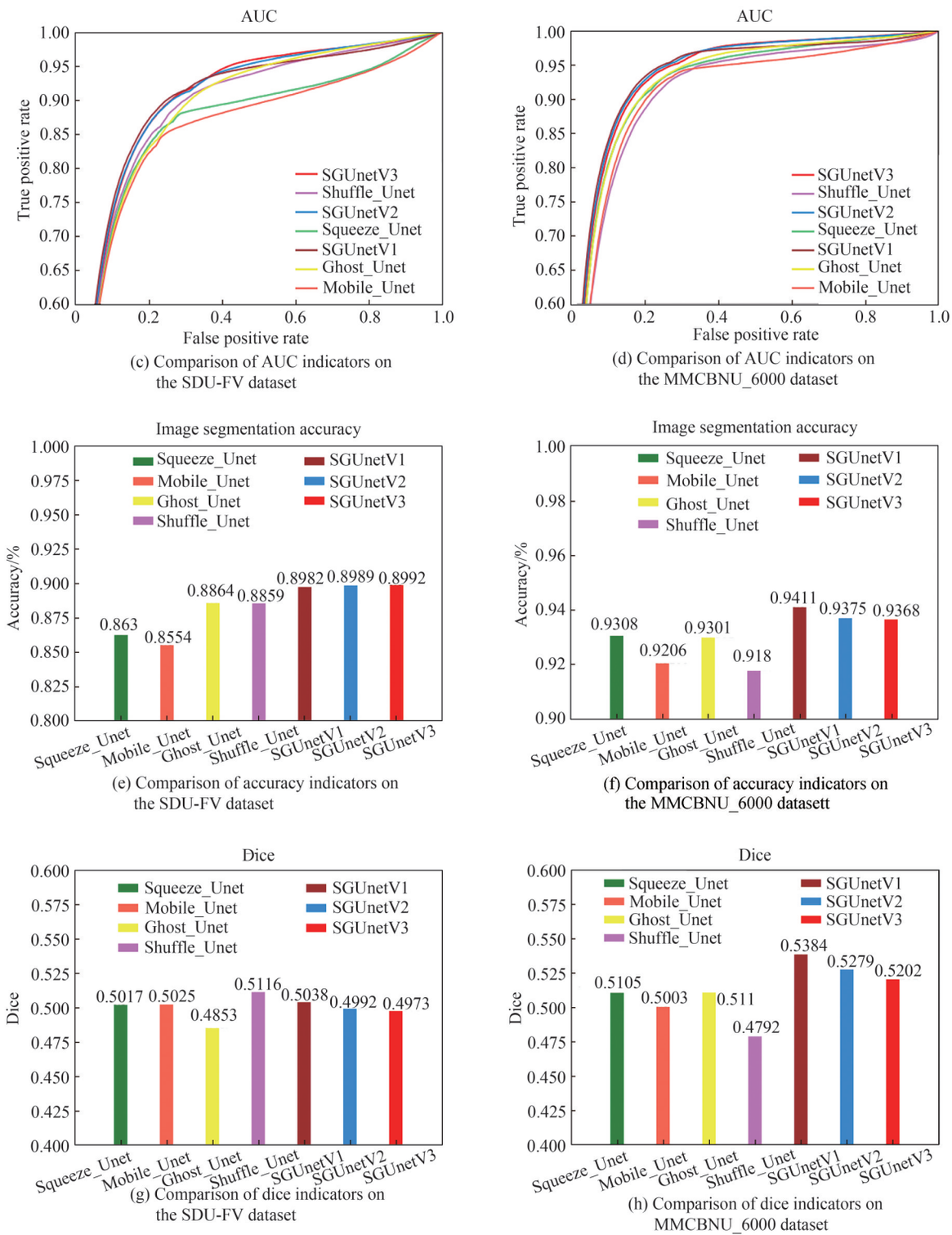


图 10　与经典轻量级网络的重要指标对比

Fig.10　Comparison of important indicators with classic lightweight networks

图 10(c)～(h)可知，SGUnetV1模型在AUC、Accuracy和Dice三个重要的分割指标上超过了以往的轻量化方法，证明本文方法在压缩模型同时兼顾分割性能提升。SGUnet系列网络不仅是在轻量化方面取得了好的成果，在性能提升方面也得到明显的增长，这一结果进一步证实了本文方法的优越性。

从上述实验结果可以说明SGUnet系列网络可以轻松地部署在计算能力较弱的平台上，本文方法不仅压缩了模型大小，减轻了运行模型所需的苛刻的硬件条件，并且取得了优秀的网络性能，分割效果和运算速度。

3 结论

针对在嵌入式平台上部署轻量级指静脉分割网络时，存在参数减小导致分割性能急剧下降、算力受限和实时性等问题，本文提出了一种超轻量级指静脉纹络实时分割网络。通过沙漏状的深度可分离卷积和轻量的ECA模块构建网络基础块，减小模型参数和提高分割性能；通过Cheap operation生成重影的特征图，减少冗余特征图；利用特征信息交互打破组间信息传递障碍，解决了分组卷积带来的难题。SGUnet系列网络在两个指静脉数据集上取得了出色的效果，在模型大小上，SGUnetV3网络参数量仅为145k，Flop仅为13M，Mult-Add仅为10M；在分割性能上，SGUnet系列网络的分割性能超过了以往的轻量化模型和经典大型分割网络。此外，在NVIDIA全系列嵌入式平台上验证了SGUnet网络的高效性和实时性，测试速度高达0.27秒/张。

参考文献

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4): 640-651.

[2] VIJAY B, ALEX K, ROBERTO C. SegNet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12):2481-2495.

[3] LIN G, MILAN A, SHEN C, et al. Refinenet: multi-path refinement networks for high-resolution semantic segmentation [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1925-1934.

[4] CHOLLET F. Xception: deep learning with depth wise separable convolutions [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251-1258.

[5] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size[J]. arXiv preprint arXiv:1602.07360, 2016.

[6] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[7] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510-4520.

[8] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: an extremely efficient convolutional neural network for mobile devices[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6848-6856.

[9] 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 (ECCV), 2018: 116-131.

[10] HAN K, WANG Y, TIAN Q, et al. Ghostnet: more features from Cheap operations[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1580-1589.

[11] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [J]. arXiv preprint arXiv: 1503.02531, 2015.

[12] HAN S, MAO H, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding[J]. arXiv preprint arXiv:1510.00149, 2015.

[13] RASTEGARI M, ORDONEZ V, REDMON J, et al. Xnor-net: Imagenet classification using binary convolutional neural networks[C]. European Conference on Computer Vision, Springer, Cham, 2016: 525-542.

[14] WEN W, XU C, WU C, et al. Coordinating filters for faster deep neural networks [C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 658-666.

[15] HOWARD A, SANLDER M, CHU G, et al. Searching for mobilenetv3[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1314-1324.

[16] TAN M, CHEN B, PANG R, et al. Mnasnet: Platform-aware neural architecture search for mobile[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 2820-2828.

[17] ZOPH B, VASUDEVAN V, SHLENS J, et al. Learning transferable architectures for scalable image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8697-8710.

[18] HU Jie, SHEN Li, SUN Gang, et al. Squeeze and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017: 99.

[19] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[20] HU J, SHEN L, ALBANIE S, et al. Gather-excite: Exploiting feature context in convolutional neural networks[J]. arXiv preprint arXiv: 1810.12348, 2018.

[21] MISRA D, NALAMADA T, ARASANIPALAI A U, et al. Rotate to attend: Convolutional triplet attention module[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 3139-3148.

[22] CHEN Y, KALANTIDIS Y, Li J, et al. A^2-nets: double attention networks[J]. arXiv preprint arXiv: 1810.11579, 2018.
[23] HUANG Z, WANG X, HUANG L, et al. Ccnet: criss-cross attention for semantic segmentation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 603-612.
[24] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[J]. arXiv preprint arXiv: 2103.02907, 2021.
[25] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2015: 234-241.
[26] DAQUAN Z, HOU Q, CHEN Y, et al. Rethinking bottleneck structure for efficient mobile network design[J]. arXiv preprint arXiv: 2007.02269, 2020.
[27] WANG Q, WU B, ZHU P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[J]. arXiv preprint arXiv: 1910.03151, 2020.
[28] ZENG J, ZHU B, HUANG Y, et al. Real-time segmentation method of lightweight network for finger vein using embedded terminal technique[J]. IEEE Access, 2020, 9: 303-316.
[29] ZHOU Haoyang, FENG Bao, QI Feifei, et al. Combining MRF energy and DCE-MRI time-domain features for breast tumors segmentation algorithm[J]. Acta Photonica Sinica, 2021, 50(6): 0610002.
周皓阳, 冯宝, 齐菲菲, 等. 结合 MRF 能量和 DCE-MRI时域特征的乳腺癌灶分割算法[J]. 光子学报, 2021, 50(6): 0610002.
[30] HUANG Hong, LV Rongfei, TAO Junli, et al. Segmentation of lung nodules in CT images using improved U-Net++[J]. Acta Photonica Sinica, 2021, 50(2): 0210001.
黄鸿, 吕容飞, 陶俊利, 等. 基于改进 U-Net++的 CT 影像肺结节分割算法[J]. 光子学报, 2021, 50(2): 0210001.

An Ultra-lightweight Real-time Segmentation Network of Finger Vein Textures

ZENG Junying[1], CHEN Yucong[1], LIN Xihua[1], QIN Chuanbo[1], WANG Yinbo[1], ZHU Jingming[1], TIAN Lianfang[2], ZHAI Yikui[1], GAN Junying[1]
(1 *Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen, Guangdong* 529020, *China*)
(2 *School of Automation Science and Engineering, South China University of Technology, Guangzhou* 510640, *China*)

Abstract: Among biometric recognition technologies, finger vein recognition has attracted the attention of many researchers because of various advantages, such as noncontact collection, living body recognition, forgery difficulty, and low cost. The finger vein extraction is the key step of finger vein recognition technology, which directly affects the accuracy of the finger vein feature extraction, matching, and recognition.

Most of the existing finger vein segmentation networks consume considerable memory and computing resources, and deploying them directly to the embedded platform is difficult. The design of lightweight deep neural network architecture is the key to solving this problem. However, most lightweight models have problems, such as sharp decline of segmentation performance, limited computing power, and real-time issue et al. To solve the above problems, this paper proposes an ultralight weight real-time segmentation network of finger vein textures-SGUnet. The SGUnet network realizes real-time finger vein texture extraction on an embedded platform, which is called finger vein segmentation. Moreover, there is a need to comprehensively consider the segmentation performance, network parameter size, and running time.

First, the encoding-decoding structure is adopted in the overall network, and the hourglass shaped deep separable volume is used to actively reduce basic model parameters to realize the preliminary lightweight of the model. The lightweight and efficient attention module is used to realize the local cross-channel interaction without dimensionality reduction, improve network segmentation performance, and solve the problem of performance degradation during model compression. The attention module uses a one-

dimensional convolution neural network to weight the channel in the operation process, while the introduced parameters of the attention module have little effect on the model's burden. Second, most convolutional neural networks have a feature graph redundancy phenomenon. These redundant feature graphs have great similarities. They can be obtained from similar feature graphs through some simple changes. To solve the problem of partial feature graph redundancy, a swap operation is used to replace some "slack" convolution cores. A similar feature map is obtained through a simple mapping transformation, which ensures the consistency of network output, reduces the part of the convolution kernel, and realizes the second step lightweight of the model. Finally, to further reduce the number of parameters of the channel convolution and the problem that each group of information in group convolution cannot flow, the characteristic information of each group is randomly disrupted and reorganized using the method of characteristic information interaction to realize the information flow between group convolution, further compress the network, and ensure the performance of the model. After the above three steps of lightweight operation, an ultralightweight real-time segmentation network of finger vein textures is finally obtained.

To verify the efficiency and real-time performance of this algorithm, two public finger vein databases are used: SDU-FV of Shandong University and MMCBNU-6000 of Quanbei National University of Korea. In the training process, four-fifths of the dataset is randomly selected as the training set and the remaining one-fifth as the test set. In the training and testing, the blocking strategy is adopted for the original image. Each image is divided into multiple patches. When the width and height are five steps, multiple continuous overlapping blocks are extracted from each image. The probability that the pixel is a vein is obtained by averaging the probability of all prediction blocks covering the pixel. To ensure that the memory limit and real-time performance of the hardware platform are not exceeded, selecting the patch with a step of five in terms of index and time is appropriate. After the network outputs the patch results, according to the order of sub patches, the overlapping sliding window strategy is adopted to retain the central region results, discard inaccurate image edges, and resplice them into a complete original image.

In the experiment, SGUnet is compared with different segmentation networks, and the comparative experiment is conducted on the embedded platform. Compared with the traditional Unet segmentation network, the parameters of SGUnet model are approximately 1%, and MultAdds are approximately 0.5% of the traditional Unet segmentation network. We verify the network performance on two public finger vein datasets: SDU-FV and MMCBNU-6000. The results show that the segmentation performance of SGUnet network is not only better than that of large segmentation networks Unet, DU-Net, and R2U-net, but also surpasses the classic lightweight models squeeze-Unet, mobile-Unet, shuffle-Unet, and Ghost-Unet, Its performance indexes accuracy, dice and AUC reach 94.11%, 0.538 4, and 0.935 4, respectively. Compared with previous work, the proposed network has made great progress, in which the final parameter is only 145K and Flops is only 13M, and it surpasses previous lightweight models. Moreover, SGUnet network meets the low computing power requirements of the embedded platform and can be easily deployed on the whole series of NVIDIA embedded platforms to realize the real-time segmentation of finger vein veins. The test speed of finger vein veins extraction is as high as 0.27 seconds/piece.

Key words: Finger vein segmentation; Lightweight network; Embedded platform; Model compression; Real-time segmentation network; Image segmentation; Convolutional neural network

OCIS Codes: 100.4996; 100.2960; 100.2000

Foundation item: National Natural Science Foundation of China (No.61771347), Special Project in Key Areas of Artificial Intelligence in Guangdong Universities (No. 2019KZDZX1017), Open Fund of Guangdong Key Laboratory of Digital Signal and Image Processing Technology(Nos. 2019GDDSIPL-03, 2020GDDSIPL-03), Special Project in Key Areas in Guangdong Universities(No.2020ZDZX3031), Guangdong Basic and Applied Basic Research Foundation (Nos. 2021A1515011576, 2019A1515010716), 2021 Jiangmen City Basic and Theoretical Scientific Research Science and Technology Plan Project (Jiangke [2021] No. 87)