计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(11):157−166 [doi: 10.15888/j.cnki.csa.008793]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进 Inception-v3 网络的手势图像识别①

邓志军, 田秋红

(浙江理工大学 信息学院, 杭州 310018)
通信作者: 田秋红, E-mail: tianqiuhong@zstu.edu.cn

摘　要: 鉴于 Inception-v3 网络参数量过大的问题, 本文提出了一种有效的手势图像识别方法, 能够满足在模型参数量较少的情况下高精度手势识别的需求. 本文利用 Inception-v3 的结构, 对原 Inception-v3 的 Inception 模块重新进行设计, 降低学习的参数量和难度, 结合残差连接, 保护信息的完整性, 防止网络退化, 引入注意力机制模块, 让模型聚焦于有用的信息而淡化无用信息, 在一定程度上也防止了模型的过拟合, 并且在模型中进行上采样与低层特征进行特征融合, 融合后的特征比原输入特征更具有判别能力, 进一步提高模型的准确率. 实验结果表明改进的 Inception-v3 网络的参数量仅为 1.65 M, 而且拥有更高的准确率和更快的收敛速度. 将 ASL 手语数据集与孟加拉手语数据集分别打乱, 然后按照 4:1 的比例单独划分出训练集和验证集. 改进的 Inception-v3 在 ASL 手语数据集与孟加拉手语数据集上的识别率分别达到了 100% 和 95.33%.

关键词: 手语识别; Inception-v3 网络; 注意力机制模块; 上采样; 特征融合; 深度学习; 卷积神经网络 (CNN)

引用格式: 邓志军,田秋红.改进 Inception-v3 网络的手势图像识别.计算机系统应用,2022,31(11):157–166. http://www.c-s-a.org.cn/1003-3254/8793.html

Improved Inception-v3 Network for Gesture Image Recognition

DENG Zhi-Jun, TIAN Qiu-Hong

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In view of a large quantity of parameters in the Inception-v3 network, this study proposes an effective gesture image recognition method, which can meet the needs of high-precision gesture recognition with few model parameters. In this study, the structure of Inception-v3 is used to redesign the Inception module of the original Inception-v3 to reduce the number and difficulty of learning parameters, and with the residual connection, the integrity of information is protected while the network degradation is prevented. The attention mechanism module is introduced to make the model focus on useful information and dilute useless information, and to a certain extent, it also prevents the overfitting of the model. Moreover, the feature fusion is carried out between the up-sampling and the low-level feature in the model, and the fused feature has better discrimination than the original input feature, which further improves the accuracy of the model. The experimental results indicate that the quantity of the parameters in the improved Inception-v3 network is only 1.65 M, and it has higher accuracy and faster convergence speed. Then, the ASL sign language dataset and the Bangladesh sign language dataset are jumbled separately, and the training set and validation set are divided at a ratio of 4:1. The recognition rates of the improved Inception-v3 on the ASL sign language dataset and Bangladesh sign language dataset are 100% and 95.33%, respectively.

Key words: sign language recognition; Inception-v3 network; attention mechanism module; up-sampling; feature fusion; deep learning; convolutional neural network (CNN)

① 基金项目: 国家自然科学基金 (51405448); 浙江理工大学博士科研启动项目 (11122932611817); 国家级大学生创新创业训练计划 (11120032382104); 浙江理工大学大学生科创项目 (11120032662023); 浙江理工大学信息学院教育教学改革研究项目 (11120033312202)
收稿时间: 2022-02-23; 修改时间: 2022-04-02; 采用时间: 2022-04-12; csa 在线出版时间: 2022-07-14

手势是一种包含信息量最多的一种人体语言, 它生动, 形象, 直观, 具有很强的视觉效果, 手势识别技术是一种建立在人与计算机之间的一种简单的人机交互方式. 如今手势识别技术已经悄然改变了我们的生活方式, 其广泛应用于游戏娱乐、计算机辅助哑语教学、医疗研究、智能设备、机器人控制等. 手势识别的研究不仅有助于改善聋哑人的生活和教育条件, 还将推动人工智能领域的发展, 所以手势识别的研究具有重要意义.

早期研究人员对手势识别的研究是基于数据手套来实现的[1]. 数据手套的每个指关节均可以通过两个传感器采集其弯曲程度的数据, 同时在两个手指之间还有一个传感器用于采集两个手指之间的角度, 将传感器的输出数据进行计算就可以得出相应的手势. 数据手套可以直接的获得手的空间三维信息和手指的运动信息, 进行实时识别, 但其手势识别网络过于依赖设备, 而且数据手套设备价格昂贵, 不易于手势识别技术的推广. 随着计算机技术的不断提高, 基于视觉的手势识别技术逐渐成为主要研究方向[2]. 相较于基于数据手套的手势识别技术, 基于视觉的手势识别技术设备简单, 只需一个或多个摄像头采集图像数据, 然后运用图像处理技术和机器学习的相关领域算法实现对手势的识别与分析. 传统的基于视觉的手势识别技术是基于人工设计的手势识别方法[3], 该方法针对特定的数据集提取适合的手势特征作为识别的特征, 通过计算机图形学对二维的手势图像属性, 如手的颜色、纹理、边界、轮廓等进行分析. 它将待识别手势的特征参数与预先存储的模板特征参数进行匹配, 学习能力并不强, 提取的手势特征单一, 容易受背景和光照等因素的影响. 为解决单一特征的手势识别在复杂背景下的识别率, 陈影柔等提出了一种基于多特征加权融合的识别方法[4]. 为追求更高的准确率, 基于神经网络的手势识别方法在手势识别领域得到广泛的应用[5–8]. 神经网络是基于统计学和概率学方法, 具有抗干扰性、自学习性、易控制性和处理高效性等特点. 基于深度学习的卷积神经网络具有层次结构化、权值共享、区域局部感应、特征提取和识别分类相结合的全局分类的特点, 能够逐层自动地学习到合适的特征并进行分类. 例如 R-CNN (区域卷积神经网络)[9]、SPP-Net 网络 (空间金字塔池化网络)[10] 和 Faster R-CNN[11] 等. 这些算法的精确度不断提高, 但是忽略了参数影响, 导致参数过大, 增加了计算的负荷.

Inception-v3 网络模型[12] 是谷歌团队提出的一种图像分类模型, 其由 Inception-v1 网络模型和 Inception-v2[13] 网络模型改进而来, 在识别精度和速度方面处于领先地位. 原始 Inception-v3 网络模型中的主干网络 Inception 模块参数量大, 模型占用内存容量大, 计算复杂度高, 对硬件的计算能力需求大. 针对上述问题, 本文提出一种改进 Inception-v3 网络的手势图像识别方法, 在保证手势识别准确率的同时极大地减少了网络模型所需的存储容量和参数量.

1 Inception-v3 模型原理

Inception 系列网络模型是由 Szegedy 等[14] 提出的. 该模型是为了解决如何在有限的计算资源内, 进一步提出网络的性能. 在当初提升网络性能最直接的方法就是增加网络的深度和宽度. 其中深度是指网络的层数, 宽度是指网络的每层的通道数. 但是这两个方法会导致两个问题: 1) 巨大的参数量导致模型容易发生过拟合. 2) 均匀地增加网络的大小, 会导致计算量的加大. 因此为解决上述问题, Szegedy 等[14] 提出了 Inception 结构, 该结构将稀疏矩阵聚类成相对密集的子矩阵, 提高了计算的性能. Inception-v3 网络模型是在 Inception-v2 网络模型的基础上提出的, 其主要思想就是分解卷积和激进的正则化. 在分解卷积中, Inception-v3 网络模型使用了两种不同的方法: 1) 将大卷积分解为小卷积, 使用了两个级联的 3×3 的滤波器来代替一个 5×5 的滤波器. 2) 将对称卷积分解为非对称卷积, 将 3×3 的卷积核分解为 1×3 和 3×1 的卷积核, 输出的特征尺寸不变但是进一步降低了计算量. 在激进的正则化中, Inception-v3 网络模型使用了 3 种不同的方法: 1) 使用批归一化, BN (batch normalization) 层. 2) Inception-v3 网络模型中的辅助分类器起到了正则化器的作用. 3) 使用标签平滑化来进行网络模型的正则化. Inception-v3 网络模型的网络图见表 1 所示.

2 改进 Inception-v3 模型设计

由表 1 可以看出, 原 Inception-v3 模型的网络层次深, 网络参数量过大, 模型的计算复杂度很高. 近年来一些经典的轻量级神经网络通过采用效率更高的计算方式来减少网络模型的参数量, 相应的可以减少参数的计算和存储, 来实现网络的优化加速, 以满足嵌入式端对于图像识别速度的要求. Mobilenet 系列[15–17] 是 Google 团队提出的一种轻量级神经网络, 其主要思想就是将原神经网络中的传统的标准的卷积替换成深度

可分离卷积, Google 团队在 Mobilenet 网络模型后续的改进中又加入倒残差结构以及线性瓶颈层结构, 构建更高效的网络模型. Shufflenet 系列模型[18,19] 主要是采用组卷积的方式减少参数和运算量. 受 Mobilenet 的深度可分离卷积的启发, 本文引入深度可分离卷积层, 对原 Inception-v3 的 Inception 模块重新进行设计, 结合残差连接, 引入注意力机制模块, 提出了一种改进的 Inception-v3 网络模型. 改进的 Inception-v3 网络模型较原模型相比拥有更少的参数量和更快的收敛速度和更高的识别准确度. 改进 Inception-v3 模型的整体结构如图 1 所示.

表 1 Inception-v3 网络结构

Type	Batch size	Input size
conv	3×3/1	299×299/3
conv	3×3/1	149×149/32
conv	3×3/1	147×147/32
pool	3×3/2	147×147/64
conv	3×3/1	73×73/64
conv	3×3/2	71×71/80
conv	3×3/1	35×35×192
3×Inception	—	35×35×288
5×Inception	—	17×17×768
2×Inception	—	8×8×1280
pool	8×8	8×8×2048
linear	logits	1×1×2048
Softmax	classifier	1×1×1000

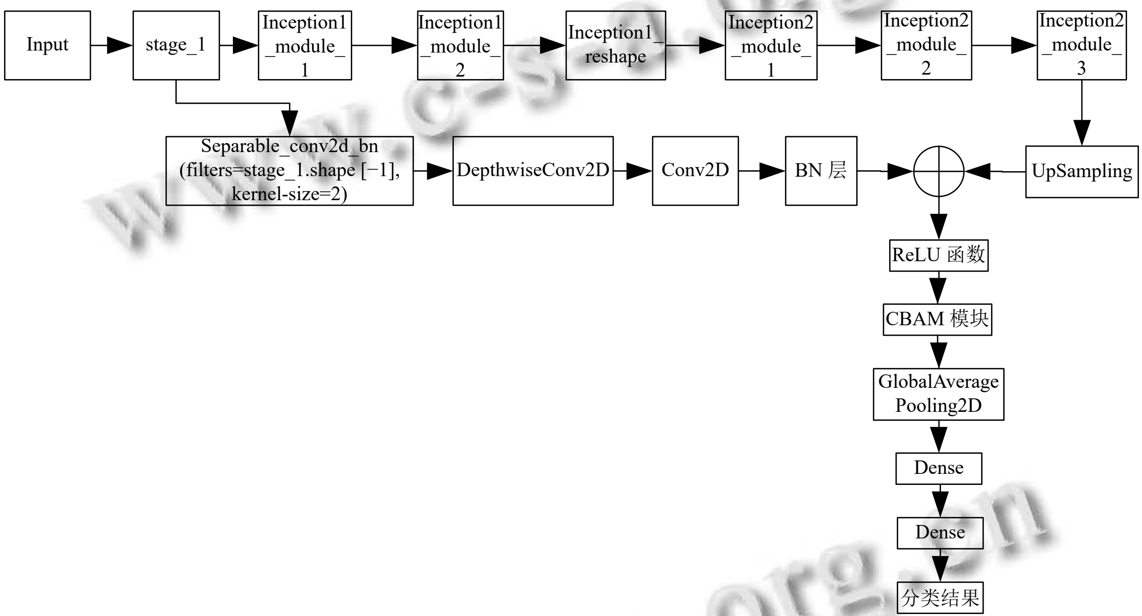

图 1 改进 Inception-v3 网络结构

将 Inception-v3 网络进行整体改进, 搭建改进后的 stage_1 网络、Inception1_module_1 网络、Inception1_module_2 网络、Inception_reshape 网络、Inception2_module_1 网络、Inception2_module_2 网络和 Inception2_module_3 网络; 每个改进后的网络均根据 Inception-v3 网络进行整体改进后直接获得. 因为原 Inception-v3 模型的 Inception 模块参数量太大, 增加了计算的复杂度, 所以本文对原 Inception-v3 模型中的 Inception 模块进行重新设计.

针对本文的改进 Inception-v3 网络的 Inception 系列网络, 即不包括 stage_1 网络, 按照式 (1)–式 (9) 设置处理:

$$C_z(x) = ReLU \cdot BN \cdot conv(x) \tag{1}$$

$$SC_z(x) = ReLU \cdot BN \cdot conv \cdot sconv(x) \tag{2}$$

$$b_{1\times1}(x) = C_1(x) \tag{3}$$

$$b_{3\times3}(x) = SC_3 \cdot C_1(x) \tag{4}$$

$$b_{5\times5}(x) = SC_5 \cdot C_1(x) \tag{5}$$

$$db_{3\times3}(x) = SC_3 \cdot SC_3 \cdot C_1(x) \tag{6}$$

$$mp_{3\times3}(x) = C_1 \cdot maxpooling_{3\times3}(x) \tag{7}$$

$$ap_{3\times3}(x) = C_1 \cdot avgpooling_{3\times3}(x) \tag{8}$$

$$md(x) = b_{1\times1}(x) \circ b_{3\times3}(x) \circ b_{5\times5}(x) \circ mp_{3\times3}s(x) \circ ap_{3\times3}(x) \tag{9}$$

其中, x表示输入; $C_z(x)$表示 z 阶普通卷积模块, z表示

卷积层的维度, 即当$z=1$时, 普通卷积模块的卷积层的卷积核大小为 1×1; ·表示组合操作; $ReLU$代表 ReLU 激活函数; BN代表 BN 层, 即标准化层; $conv(x)$表示卷积层; $SC_z(x)$表示 z 阶可分离卷积模块; $conv$表示由上一步的输出作为输入的卷积层; $sconv(x)$表示可分离卷积层; $b_{1\times1}(x)$表示由一个一阶普通卷积模块构成的网络; $C_1(x)$ 表示一阶普通卷积模块; $b_{3\times3}(x)$表示由一个一阶普通卷积模块和一个卷积核大小为 3×3 的可分离卷积模块构成的网络; SC_3表示由上一步的输出作为输入的三阶可分离卷积模块; $b_{5\times5}(x)$表示由一个一阶普通卷积模块和一个卷积核大小为 5×5 的可分离卷积模块构成的网络; SC_5表示由上一步的输出作为输入的五阶可分离卷积模块; $db_{3\times3}(x)$表示由一个一阶普通卷积模块和两个卷积核大小为 3×3 的可分离卷积模块构成的网络; $mp_{3\times3}(x)$表示由一个一阶普通卷积模块和一个卷积核大小为 3×3 的最大池化层构成的网络; C_1表示由上一步的输出作为输入的一阶普通卷积模块; $maxpooling_{3\times3}(x)$表示卷积核大小为 3×3 的最大池化层; $ap_{3\times3}(x)$表示由一个一阶普通卷积模块和一个卷积核大小为 3×3 的平均池化层构成的网络; $avgpooling_{3\times3}(x)$表示卷积核大小为 3×3 的平均池化层; $md(x)$表示一个网络模块, 其中$\circ$表示拼接; 它由$b_{1\times1}(x)$、$b_{3\times3}(x)$、$b_{5\times5}(x)$、$mp_{3\times3}(x)$以及$ap_{3\times3}(x)$的输出依次在特征向量的通道数方向拼接组成.

2.1 改进 Inception-v3 模型中的 stage_1 模块

图 1 中的 stage_1 网络结构包括 4 个三阶普通卷积模块, 一个一阶普通卷积模块, 两个最大池化层, 一个 CBAM 注意力机制模块, 如图 2 所示. 其中普通卷积模块 conv2d_bn 中一阶普通卷积模块和三阶普通卷积模块的结构相同, 只是模块中卷积层的卷积核分别是 1×1 和 3×3. 所述的一阶普通卷积模块和三阶普通卷积模块的网络结构相同, 具体为一个 Conv2D 卷积层、一个 BN 层、一个 ReLU 激活函数依次连接组成. 一阶普通卷积模块的 Conv2D 卷积层的卷积核大小为 1×1, 三阶普通卷积模块的 Conv2D 卷积层的卷积核大小为 3×3. 本文的普通卷积模块 conv2d_bn 的简化表达式为:

$$\text{conv2d\_bn}(\text{filters}=x_i, \text{kernel-size}=x_o)$$

其中, filters 为 conv2d_bn 模块卷积层过滤器的个数, kernel-size 为过滤器的大小.

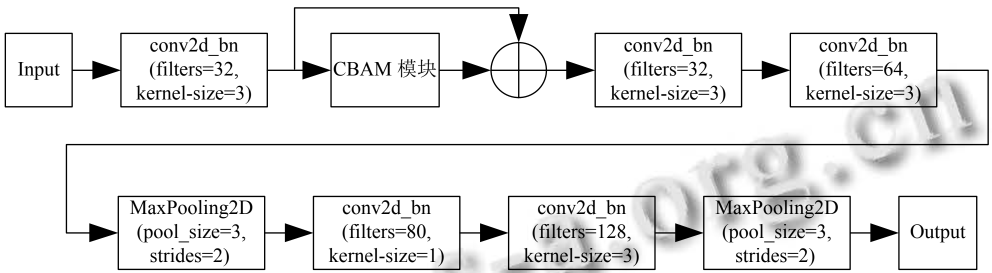

图 2　Stage_1 网络结构

本文最大池化层 MaxPooling2D, 简化表达式为:

$$\text{MaxPooling2D}(\text{pool\_size}=x, \text{strides}=s)$$

其中, pool_size 为池化层的池化窗口大小, strides 为池化的步长.

本文在 stage_1 网络结构拓扑中加入了 CBAM 注意力机制模块[20], 一种用于前馈卷积神经网络的简单而有效的注意力模块. CBAM 注意力机制模块包含 2 个独立的子模块, 通道注意力模块 (channel attention module, CAM) 和空间注意力模块 (spatial attention module, SAM), 这样不仅能够节约参数和计算力, 而且保证了其能够做为即插即用的模块集成到网络架构中. CBAM 注意力机制模块的输入先经通道注意力机制模块处理, 后经空间注意力机制模块处理, 然后将处理过的特征图与原输入特征图相乘以进行自适应特征优化, 增强了网络对空间和通道的关注度, 提高模型的精度.

2.2 改进 Inception-v3 模型中的 Inception 模块

所述的 Inception1_module_1 网络结构包括一个五阶可分离卷积模块, 一个三阶可分离卷积模块, 5 个一阶普通卷积模块, 一个平均池化层, 一个最大池化层, 一个拼接层, 一个 BN 层, 如图 3 所示. 其中可分离卷积模块 separate_conv2d_bn 结构主要由一个深度可分

离卷积层、一个 Conv2D 卷积层、一个 BN 层和 ReLU 激活函数依次连接构成, 由三阶可分离卷积模块的输入或五阶可分离卷积模块的输入依次输入到深度可分离卷积层、Conv2D 卷积层、BN 层和 ReLU 激活函数处理并输出; 三阶可分离卷积模块的深度可分离卷积层的卷积核大小为 3×3, 五阶可分离卷积模块的深度可分离卷积层的卷积核大小为 5×5. 本文的可分离卷积模块 separate_conv2d_bn 的简化表达式为:

$$\text{separable\_conv2d\_bn}(\text{filters} = x_i, \text{kernel-size} = x_o)$$

其中, filters 为 separable_conv2d_bn 模块中 Conv2D 卷积层过滤器的个数, kernel-size 为深度可分离卷积层中过滤器的大小, Conv2D 卷积层中过滤器的大小为 1, 步长为 1. Conv2D 卷积层的作用是将深度可分离卷积层输出特征中的通道数关联起来.

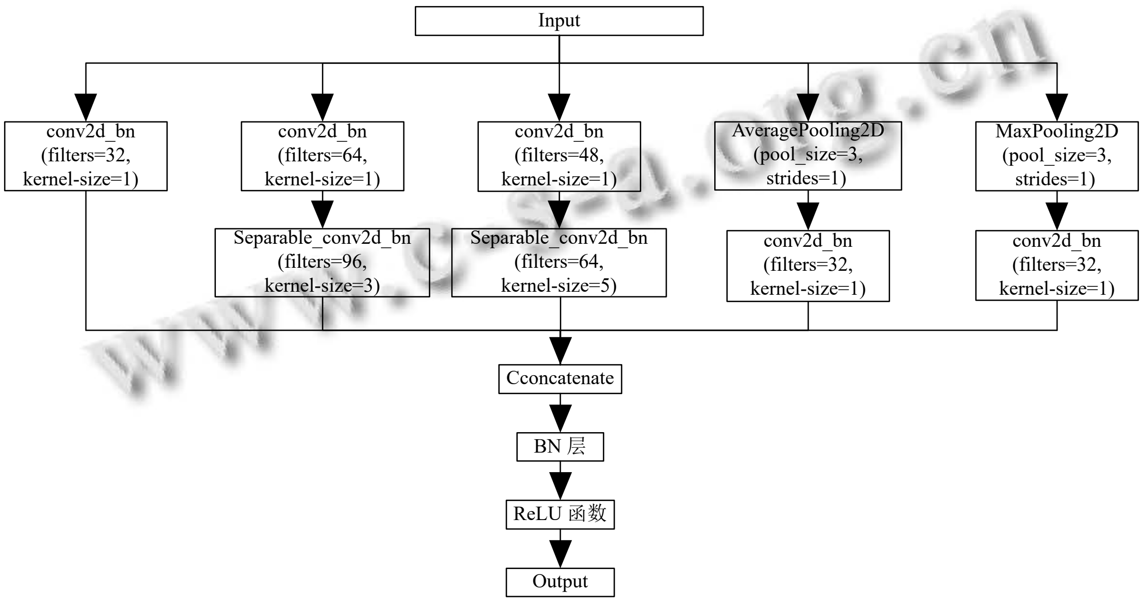

图 3 Inception1_module_1 网络结构

本文的平均池化层 AveragePooling2D, 简化表达式为:

$$\text{AveragePooling2D}(\text{pool\_size} = x, \text{strides} = s)$$

其中, pool_size 为池化层的池化窗口大小, strides 为池化的步长.

拼接层是将输入到拼接层的特征向量在通道数上进行拼接, 使得输出的特征向量包含不同的感受野. Inception1_module_1 网络的公式如式 (10) 所示:

$$I_{1\_1}(x) = ReLU \cdot BN \cdot md(x) \tag{10}$$

所述 Inception1_module_2 网络结构与 Inception1_module_1 网络结构相似, 均包括一个五阶可分离卷积模块, 一个三阶可分离卷积模块, 5 个一阶普通卷积模块, 一个平均池化层, 一个最大池化层, 一个拼接层, 一个 BN 层, 如图 4 所示. 与 Inception1_module_1 网络结构不同的是, Inception1_module_2 采用了 ResNet[21] 的残差连接的思想. 特征向量并联输入到拼接层在通道数上进行特征向量的拼接后, 拼接层输出的特征向量与 Inception1_module_2 网络结构的原输入特征向量进行相加再输入到 BN 层和 ReLU 激活函数得到 Inception1_module_2 的输出. 残差连接的使用可以在一定程度上防止网络的退化以及过拟合问题. Inception1_module_2 网络的公式如式 (11) 所示. 其中⊕表示残差连接.

$$I_{1\_2}(x) = ReLU \cdot BN \cdot [x \oplus md(x)] \tag{11}$$

所述的 Inception2_module_1 网络结构包括 3 个三阶可分离卷积模块, 5 个一阶普通卷积模块, 一个平均池化层, 一个最大池化层, 一个拼接层, 一个 BN 层, 如图 5 所示. Inception2_module_1 网络与 Inception1 系列网络不同的是为了减少参数量, 增加非线性因素, 将 5×5 的可分离卷积模块改为两个连续的 3×3 的可分离卷积模块, 所得到的第 5 特征向量组尺寸与原可分离卷积模块相同; 并增大了两个连续 3×3 的可分离卷积模块的比例. Inception2_module_1 网络的公式如式

(12) 所示:

$$I_{2\_1}(x) = ReLU \cdot BN \cdot \begin{bmatrix} b_{1\times1}(x) \circ b_{3\times3}(x) \circ \\ db_{3\times3}(x) \circ mp_{3\times3}(x) \circ ap_{3\times3}(x) \end{bmatrix} \tag{12}$$

所述的 Inception2_module_2 和 Inception2-module_3 的网络结构相同. 与 Inception2_module_1 相同, 均包括 3 个三阶可分离卷积模块, 5 个一阶普通卷积模块, 一个平均池化层, 一个最大池化层, 一个拼接层, 一个 BN 层. 不同的是 Inception2_module_2 和 Inception2_module_3 也采用了残差连接的思想, 防止网络模型在层数加深的情况下退化. Inception2_module_3 网络目的是加深网络模型的深度, 提高模型的准确度.

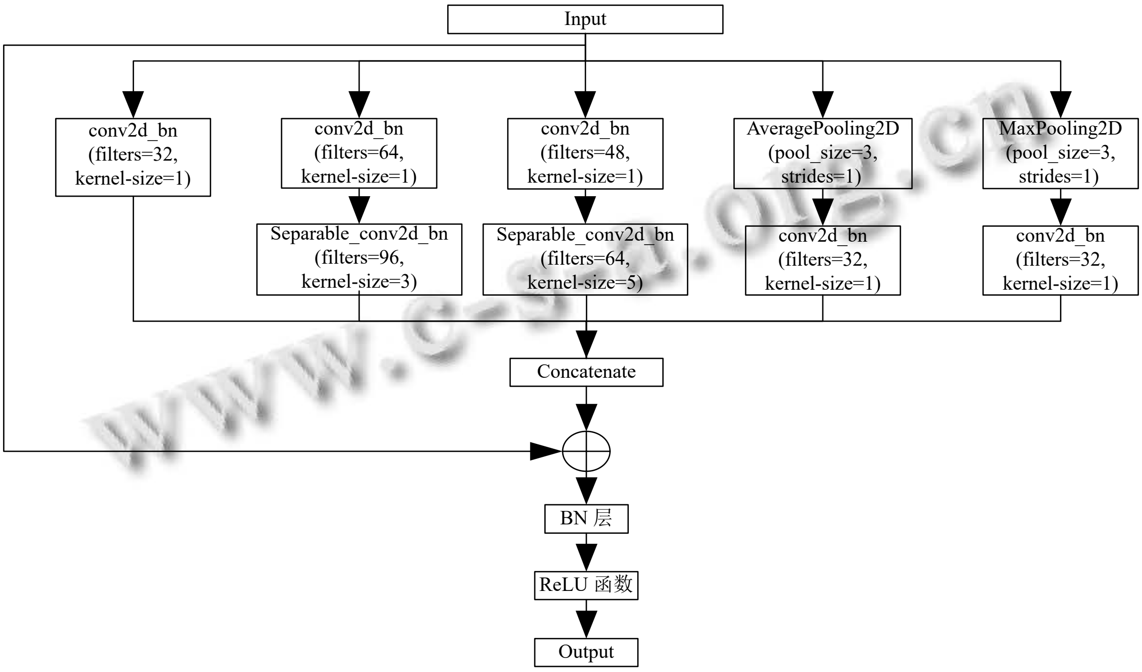

图 4　Inception1_module_2 网络结构

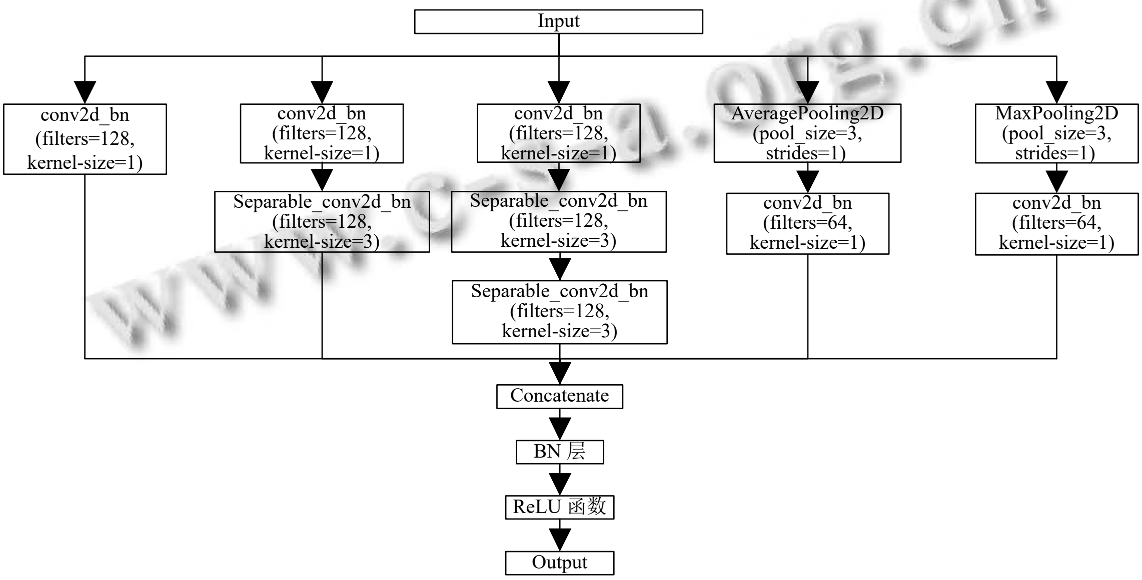

图 5　Inception2_module_1 网络结构

2.3　改进 Inception-v3 模型的 Inception_reshape 模块

Inception_reshape 网络结构包括一个三阶可分离卷积模块, 3 个一阶普通卷积模块, 一个平均池化层, 一个最大池化层, 一个拼接层, 一个 BN 层, 如图 6 所示.

Inception_reshape 模块的作用是将特征向量的尺寸减小, 通过 3 种方式: ① 采用三阶可分离卷积模块减小尺寸; ② 采用平均池化层减小尺寸; ③ 采用最大池化层减小尺寸. Inception_reshape 模块将这 3 种减小特征向量尺寸的方式通过拼接层在特征向量的通道数上进行拼接, 让网络模型自己训练出最合适的减小特征向量尺寸的方法. Inception_reshape 网络的公式如式 (13) 所示:

$$I_{\mathrm{reshape}}(x) = SC_{3\times3}(x) \circ mp_{3\times3}(x) \circ ap_{3\times3}(x) \tag{13}$$

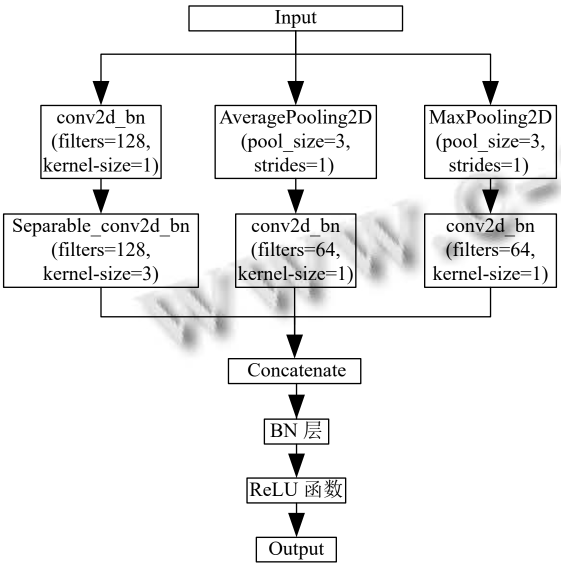

图 6 Inception_reshape 网络结构

2.4 改进 Inception-v3 模型的上采样和多尺度特征融合

图 1 中 UpSampling 模块的网络结构主要由一个上采样层、一个卷积核大小为 3×3 的 DepthwiseConv2D 深度可分离卷积层、一个卷积核大小为 1×1 的 Conv2D 卷积层和一个 BN 层依次连接构成, 作用是对 Inception2_module3 输出的特征向量进行上采样, 将特征向量的尺寸翻倍. 同时本文对 stage_1 模块输出的特征向量依次经过一个深度可分离卷积层的卷积核大小为 2×2 的二阶可分离卷积模块, 一个卷积核大小为 3×3 的 DepthwiseConv2D 深度可分离卷积层, 一个卷积核大小为 1×1 的 Conv2D 卷积层和一个 BN 层, 进行卷积处理后, 得到尺寸大小与 UpSampling 输出的特征向量尺寸大小相同的特征向量, 再于 UpSampling 模块输出的特征向量进行特征融合. 进行低层特征向量与高层特征向量融合的原因是低层特征分辨率更高, 包含更多位置、细节信息, 但是由于经过的卷积更少, 其语义性更低, 噪声更多. 高层特征具有更强的语义信息, 但是分辨率很低, 对细节的感知能力较差. 将两者进行融合合并成一个比输入特征更具有判别能力的特征可以提高模型的准确率.

3 数据集的介绍与实验结果分析

3.1 数据集介绍

本发明采用 ASL (American sign language) 开源手语数据集, 采集的原始图像数据集中包含了不同角度、不同光照、不同大小以及不同背景环境下的手势图像, 如图 7 所示, 共包含 28 种手势类别以及非手势类别, 共 29 个分类类别, 手势图像总数为 87 000 张. ASL 手语数据集的手势图像背景区域较为干净, 背景颜色与手的肤色差别较大, 包含了手势在不同角度受不同光照强度下的手势图片. 本实验将手势图像打乱后按照 4:1 的比例单独划分出训练集和验证集进行模型的训练和验证.

图 7 ASL 数据集

3.2 改进 Inception-v3 网络模型参数量与时间复杂度对比

本文方法参数量与原 Inception-v3 网络模型参数量对比见表 2. 由表 2 可见, 原 Inception-v3 网络的参数量为 21 862 205, 而本文方法的参数量为 1 647 637. 本文方法的参数量仅约为原模型的 0.075 倍, 参数量减少了约 13.3 倍, 极大地降低了模型的参数量. 同时原 Inception-v3 网络的时间复杂度 GFLOPs 为 11.45, 而本文方法的时间复杂度 GFLOPs 为 2.03. 本文方法的 GFLOPs 仅为原模型的 0.177 倍, 减少了约 5.64 倍. 对比原 Inception-v3 网络与本文方法的时间复杂度可知, 本文方法减轻了模型的计算复杂度, 加快了模型的训练速度.

表 2 原 Inception-v3 模型与本文方法参数量对比

对比指标	原Inception-v3	本文方法
参数量	21 862 205	1 647 637
GFLOPs	11.45	2.03

3.3 消融实验

为了验证本文的改进 Inception-v3 网络模型中的 CBAM 注意力机制模块与多尺度特征融合对改进模型识别准确率的影响, 本文设计了 3 组网络模型, 分别为包含了 CBAM 注意力机制模块与多尺度特征融合的本文的改进 Inception-v3 模型; 不包含 CBAM 注意力机制模块的改进的 Inception-v3 模型; 不包含多尺度特征融合的改进的 Inception-v3 模型. 将这 3 组网络模型在孟加拉数据集中采用相同的批处理尺寸、学习率和迭代次数进行训练, 实验结果如表 3 所示. 本文的改进 Inception-v3 网络模型的识别准确率为 95.33%, 不包含 CBAM 模块的改进的 Inception-v3 模型的识别准确度为 90.92%, 不包含多尺度特征融合的改进的 Inception-v3 模型的识别准确率为 93.03%. 添加 CBAM 注意力机制模块和多尺度特征融合可以分别提高网络模型识别准确率 4.41% 与 2.3%. 由此可见, CBAM 注意力机制模块有助于网络模型更好地关注有效特征, 同时多尺度特征融合使得特征更具有判别能力, 提高模型的提取特征的能力, 从而提高网络模型的识别准确率.

表 3 孟加拉手语数据集中本文方法与不含 CBAM, 不含特征融合的识别准确率 (%)

模型	本文方法	不含CBAM	不含特征融合
准确率	95.33	90.92	93.03

3.4 网络结构对比分析

本实验将 batch_size 设置为 32, 同时为防止手势图像识别模型出现过拟合问题, 将最大迭代次数设为 20 次. 将网络的初始学习率 lr 设为 0.001, 并且通过回调函数在模型训练过程中对学习率进行实时更新, 衰减系数设为 0.9. 当模型的验证集准确率下降时, 学习率减小的动作就会被触发, 使得学习率的大小减小为原来大小的 0.9 倍.

为了验证本文的改进 Inception-v3 网络模型的手势图像识别的可行性, 本文在 ASL 手语数据集中采用相同的批处理尺寸、学习率和迭代次数构建了 4 组网络模型进行对比试验, 分别为 VGG16 网络模型[22] 网络模型、ResNet50 网络模型[23]、Inception-v3 网络模型、以及本文改进的 Inception-v3 网络模型. 实验结果如图 8 所示. 从图 8 可以看出, 本文的改进 Inception-v3 网络模型相比于其他 3 种模型的识别准确率提升过程更加稳定且收敛速度也较快. 本文方法在训练至第 8 个 Epoch 时准确率趋于稳定, VGG16 网络、ResNet50 网络以及原 Inception-v3 网络分别在第 18 个 Epoch、第 11 个 Epoch 以及第 7 个 Epoch 时趋于稳定, 且本文方法的准确率在训练过程中上下波动幅度均小于其他 3 种模型. 本文在 ASL 手语数据集打乱后按照 4:1 比例单独划分出训练集和验证集测试 4 种模型, 当模型的识别准确率稳定后, 本文的改进 Inception-v3 网络模型的准确率为 100%, VGG16 网络模型的准确率为 99.95%, ResNet50 网络模型的准确率为 99.99%, 原 Inception-v3 网络模型的准确率为 100%. 通过比较 4 种模型在识别准确率折线趋于稳定后的准确率可以看出: 本文改进的 Inception-v3 网络模型在减少模型参数量的同时模型的稳定性、收敛速度和图像识别准确率均具有一定的提升. 对比表 4 现有手势识别模型与本文的改进 Inception-v3 网络模型的参数量, 本文的改进 Inception-v3 网络模型的参数量与传统 CNN-L4 模型相比减少了 0.05 M, 与 Capsule-EM 模型相比增加了 0.05 M, 与 ICapsule-EM 模型相比减少了 0.05 M, 与 LeNet-5 模型相比增加了 0.95 M, 与 Inception-v3 模型相比减少了 26.51 M. 对比表 4 现有手势识别模型与本文的改进 Inception-v3 网络模型在 ASL 手语数据集的识别准确率, 本文的改进 Inception-v3 网络模型的识别准确率与传统 CNN-L4 模型相比提升了 9.41%, 与 Capsule-EM 模型相比提升了 9.01%, 与 ICapsule-EM 模型相比提升了 5.44%, 与 LeNet-5 模型相比提升了 1.7%, 与 Inception-v3 模型相同. 本文的改进 Inception-v3 网络模型在模型的参数量与表 4 给出的除 Inception-v3 模型外的其他 4 种手势识别模型的参数量相差不大的情况下, 手势识别准确率均有较大提升. 在于原 Inception-v3 模型识别准确率相同的情况下, 本文改进的 Inception-v3 网络模型的参数量远小于原 Inception-v3 网络模型.

为了进一步验证本文的改进 Inception-v3 网络模型的有效性, 本文在孟加拉手语数据集 (Bengali sign language dataset) 上再一次进行了研究. 孟加拉手语数据集部分数据如图 9 所示. 该数据集包含 38 个分类类别, 训练数据集有 11 061 张, 验证数据集有 1 520 张, 手势图像总数为 12 581 张. 孟加拉手语数据集与 ASL 手语数据集相比, 手势图像的背景区域的干扰信息多, 背景颜色复杂, 包含了与肤色十分相似的颜色, 而且有些手势图像的背景区域中还包含了手势的投影, 手势

图像中的肤色也不止一种, 极大地加大了网络模型对手势识别的难度. 本文在孟加拉手语数据集中采用相同的批处理尺寸、学习率和迭代次数训练 VGG16 网络模型、ResNet50 网络模型、原 Inception-v3 网络模型以及本文改进 Inception-v3 网络模型. 其中 VGG16、ResNet50 与 Inception-v3 均采用了在 ImageNet 数据集上预先训练好的模型. 实验结果见图 10 和表 5. 从图 10 4 种网络模型的稳定性来看, 本文的改进 Inception-v3 网络模型的稳定性较好, 波动的幅度小于 ResNet50 与 Inception-v3 网络模型, 与 VGG16 的波动幅度相近. 从图 10 四种网络模型的手势识别准确率的收敛速度来看, 本文的改进 Inception-v3 网络模型的收敛速度也是快于其他 3 种网络模型. 本文方法在训练至第 11 个 Epoch 时准确率趋于稳定, VGG16 网络、ResNet50 网络以及原 Inception-v3 网络分别在第 16 个 Epoch、第 17 个 Epoch 以及第 17 个 Epoch 时趋于稳定. 由表 5 可知, VGG16 网络模型在该数据集上得到的识别准确率为 95.13%, ResNet50 网络模型在该数据集上得到的识别准确率为 89.08%, 原 Inception-v3 网络模型在该数据集上得到的识别准确率为 91.44%, 本文的改进 Inception-v3 网络模型在该数据集上得到的识别准确率为 95.33%. 与 VGG16 网络模型和 ResNet50 网络模型相比, 本文的改进 Inception-v3 网络模型的准确率分别提升了 0.2% 和 5.85%. 与原 Inception-v3 网络模型相比, 识别准确率提高了 3.89%. 对比表 5 的现有手势识别模型与本文的改进 Inception-v3 网络模型的参数量, 本文的改进 Inception-v3 网络模型的参数量与 VGG16 网络模型相比减少了 39.45 M, 与 ResNet50 网络模型相比减少了 13.81 M, 与原 Inception-v3 网络模型相比减少了 20.23 M. 由此可见, 本文的改进 Inception-v3 网络模型不仅参数量少, 而且识别的准确率也比表 5 给出的其他 3 种网络模型的准确率高.

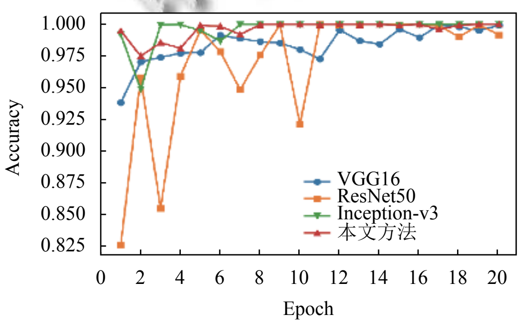

图 8　ASL 数据集中 VGG16, ResNet50, Inception-v3, 本文方法识别准确率折线图

表 4　现有手势识别模型与本文方法在 ASL 手语数据集上的检测结果

模型	准确率 (%)	模型参数量 (M)
传统CNN-L4	90.59	1.7
Capsule-EM	90.99	1.6
ICapsule-EM	94.56	1.7
LeNet-5	98.30	0.7
Inception-v3	100	28.16
本文方法	100	1.65

图 9　孟加拉手语数据集

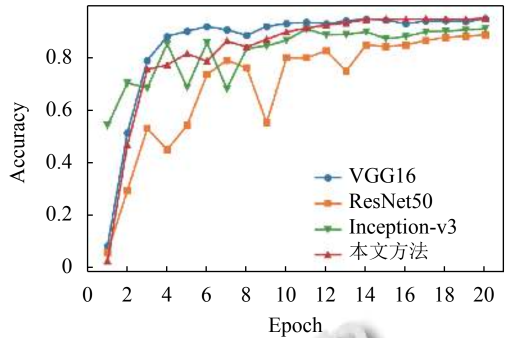

图 10　孟加拉手语数据集中 VGG16, ResNet50, Inception-v3, 本文方法识别准确率折线图

表 5　现有手势识别模型与本文方法在孟加拉手语数据集上的检测结果

模型	准确率 (%)	模型参数量 (M)
VGG16	95.13	41.1
ResNet50	89.08	15.46
Inception-v3	91.44	21.88
本文方法	95.33	1.65

4　结论

本文提出了一种改进 Inception-v3 网络的手势图像识别方法. 并对原 Inception-v3 网络模型进行了分析, 本文提出的改进 Inception-v3 网络模型对原 Inception-v3 网络模型的不足之处各方面进行了改进, 受 Mobilenet 系列网络模型的深度可分离卷积的启发, 首先对原 Inception-v3 的 Inception 模块进行重新设计, 以此减少网络模型的参数量, 然后再引入残差连接、CBAM 注意力机制模型和上采样进行特征融合来保证

模型识别的准确率和增强网络模型的适应性. 通过进行消融实验本文证明了 CBAM 注意力机制模块与多尺度特征融合是有助于提高网络模型的识别准确率. 最后本文的改进 Inception-v3 网络模型分别在 ASL 数据集和孟加拉手语数据集上进行了实验, 识别准确率分别为 100% 和 95.33%.

参考文献

1 李清水, 方志刚, 沈模卫, 等. 手势识别技术及其在人机交互中的应用. 人类工效学, 2002, 8(1): 27–29, 33.

2 田秋红, 杨慧敏, 梁庆龙, 等. 视觉动态手势识别综述. 浙江理工大学学报 (自然科学版), 2020, 43(4): 557–569.

3 Hrúz M, Trojanová J, Železný M. Local binary pattern based features for sign language recognition. Pattern Recognition and Image Analysis, 2012, 22(4): 519–526.

4 陈影柔, 田秋红, 杨慧敏, 等. 基于多特征加权融合的静态手势识别. 计算机系统应用, 2021, 30(2): 20–27. [doi: 10.1134/S1054661812040062]

5 Pigou L, Dieleman S, Kindermans PJ, *et al*. Sign language recognition using convolutional neural networks. Proceedings of the Workshop at the European Conference on Computer Vision. Zurich: Springer, 2014. 572–578.

6 Jie H, Zhou WG, Li HQ, *et al*. Sign language recognition using 3D convolutional neural networks. Proceedings of 2015 IEEE International Conference on Multimedia and Expo (ICME). Turin: IEEE, 2015. 1–6.

7 Hore S, Chatterjee S, Santhi V, *et al*. Indian sign language recognition using optimized neural networks. Proceedings of the 2015 International Conference on Information Technology and Intelligent Transportation Systems. Xi'an: Springer, 2017. 553–563.

8 Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1725–1732.

9 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

10 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(9): 1904–1916.

11 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

12 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826.

13 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: ACM, 2015. 448–456.

14 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1–9.

15 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861, 2017.

16 Sandler M, Howard A, Zhu ML, *et al*. MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4510–4520.

17 Howard A, Sandler M, Chu G, *et al*. Searching for MobileNetV3. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1314–1324.

18 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856.

19 Ma NN, Zhang XY, Zheng HT, *et al*. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 116–131.

20 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 3–19.

21 He KM, Zhang XY, Ren SQ, *et al*. Identity mappings in deep residual networks. Proceedings of the 14th European Conference on Computer Vision. Munich: Springer, 2016. 630–645.

22 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

23 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

(校对责编: 孙君艳)