DOI:10.19651/j.cnki.emt.2108324

基于轻量级卷积神经网络的手势识别检测*

牛雅睿[1] 武 一[1,2] 孙 昆[1] 卢 昊[1] 赵 普[1]
(1.河北工业大学电子信息工程学院 天津 300401;2.河北工业大学电子与通信工程国家级实验教学示范中心 天津 300401)

摘 要: 针对基于深度学习的手势识别模型参数量大、训练速度缓慢且对设备要求高,增加了成本的问题,提出了一种基于轻量级卷积神经网络的手势识别检测算法。首先利用 Ghost 模块设计轻量级主干特征提取网络,减少网络的参数量和计算量;通过引入加权双向特征金字塔网络改进特征融合网络,提升网络检测精度;最后使用 CIoU 损失函数作为边界框回归损失函数并加入 Mosaic 数据增强技术,加快模型收敛速度提升网络的鲁棒性。实验结果表明,改进后的模型大小仅为 17.9 MB,较原 YOLOv3 模型大小减小了 92.4%,平均精确度提高了 0.6%。因此新的检测方法在减少模型参数量的同时,还可保证模型的检测精度和效率,为手势识别检测提供理论参考。

关键词: 手势识别;轻量级网络;YOLOv3;Ghost 模块;加权双向特征金字塔;CIoU 损失函数
中图分类号: TP391.41 **文献标识码:** A **国家标准学科分类代码:** 510.4

Gesture recognition and detection based on lightweight convolutional neural network

Niu Yarui[1] Wu Yi[1,2] Sun Kun[1] Lu Hao[1] Zhao Pu[1]
(1. School of Electronics Information Engineering, Hebei University of Technology, Tianjin 300401, China;
2. Electronics and Communication Engineering National Experimental Teaching Demonstration Center, Hebei University of Technology, Tianjin 300401, China)

Abstract: Aiming at the problems of deep learning-based gesture recognition model with large parameters, slow training speed and high equipment requirements, which increase the cost, a gesture recognition and detection algorithm based on lightweight convolutional neural network is proposed. First, use the Ghost module to design a lightweight backbone feature extraction network to reduce the amount of parameters and calculations of the network. Improve the feature fusion network by introducing a weighted two-way feature pyramid network to improve the network detection accuracy. Finally use the CIoU loss function as the bounding box regression loss function and add Mosaic data enhancement technology to speed up model convergence and improve the robustness of the network. Experimental results show that the size of the improved model is only 17.9 MB, which is 92.4% smaller than the original YOLOv3 model, and the average accuracy is increased by 0.6%. Therefore, the new detection method can not only reduce the amount of model parameters, but also ensure the accuracy and efficiency of the model, providing a theoretical reference for gesture recognition and detection.

Keywords: gesture recognition; lightweight network; YOLOv3; Ghost module; weighted bidirectional feature pyramid; CIoU loss function

0 引 言

随着科学技术的发展,计算机正在以惊人的速度进入人类社会的各个角落,人机交互(human computer interaction,HCI)已经成为人们日常生活的重要组成部分。传统的人机交互的方式主要利用鼠标、键盘、触摸屏和数据手套等穿戴设备,但是这种接触式方法不能满足自然的交互需求,随着语音识别、手势识别等相关技术的逐渐成熟,非接触式的交互手段逐渐成为了当下研究的热门方向之一[1]。手势作为人际交往中重要的一部分,具有直观形象和易理解的特点,同时它起着加强语言的力量、丰富语言的色彩等补充和说明的作用,通过手势识别来实现人机交互

收稿日期:2021-11-12
* 基金项目:国家自然科学基金(51977059)、河北省自然科学基金(E2020202042)项目资助

的技术可以很自然地应用于手语识别、车载设备控制、智能家居等方面，是未来人机自然交互的重点研究方向之一。

根据手势识别使用的设备是否与身体接触，手势识别方法主要分为两种：1）基于可穿戴设备的方法。利用数据手套上各种传感器设备进行数据采集，将采集到的信息传输给计算机进行手势识别，但由于硬件成本昂贵和使用起来并不方便，其在常用领域中推广价值不高。2）基于计算机视觉的方法。仅使用 RGB 或深度摄像头就可以对手势进行识别，且识别的精度和速度都比较理想，符合自然的人机交互方式。

常见的基于人工特征提取的手势识别方法有模板匹配法、隐马尔可夫模型法和支持向量机法。文献[2]针对通常采取单一特征进行手势分类，但是单一特征不能代表整个图像的问题，提出一种多特征融合的手势识别方法，分别提取了梯度方向直方图（HOG）和局部二值模式（LBP）两种特征，并进行特征融合，然后将融合特征向量输入 SVM 分类器完成手势识别，该方法只能识别较为简单的手势，对具有复杂特征的手势识别效果不好，准确率仅为 62.2%；文献[3]提出一种融合静态手势特征和手部运动轨迹特征的手势交互方法，但该方法需要使用深度摄像头和 Kinect 摄像头分别提取图像信息和手部骨骼节点，手势特征提取比较繁琐；文献[4]提出一种基于多特征融合及生物启发式遗传算法优化多分类支持向量分类器的静态手势识别方法；文献[5]为了解决现有手势识别易受背景噪声干扰和算法较为复杂的问题，提出一种基于 3D 视觉的数字手势语义识别方法，但该方法只能对 0～9 数字手势识别且不能识别更多形态的手势。整体上来说，传统的手势识别方法需要人工设计特征提取方法，对具有复杂特征的手势提取比较困难，在识别准确率、识别速度和鲁棒性等方面还有诸多问题有待解决。

近年来，深度学习在目标检测、图像分类等领域得到了巨大的发展。如 YOLO[6-8]、SSD[9]、R-CNN[10] 和 Faster R-CNN[11] 等算法在目标检测和分类问题中取得了较高的准确率。文献[12]提出了一种多尺度卷积特征融合的 SSD 手势识别方法，引入了不同卷积层的特征融合思想，经过空洞卷积下采样操作与反卷积上采样操作，实现网络结构中的浅层视觉卷积层与深层语义卷积层的融合，代替原有的卷积层用于手势识别，以提高模型对中小目标手势的识别精度；文献[13]为了提高手势识别的准确性、鲁棒性以及收敛速度，提出一种基于改进残差网络和动态调整学习率的手势识别方法研究；文献[14]提出基于改进 YOLOv3 网络与贝叶斯分类器相结合的手势识别深度学习模型，虽然解决了数据易受影响问题并且增强了网络不变性，但是该方法使用的手势数据集较为简单且背景较为单一；文献[15]针对使用 YOLOv4 算法识别手势的误检和漏检较多及手势数据较少的问题，提出了一种基于改进 YOLOv4 的手势交互算法，该方法能识别复杂场景下的手势，并且能够满足实时要求。

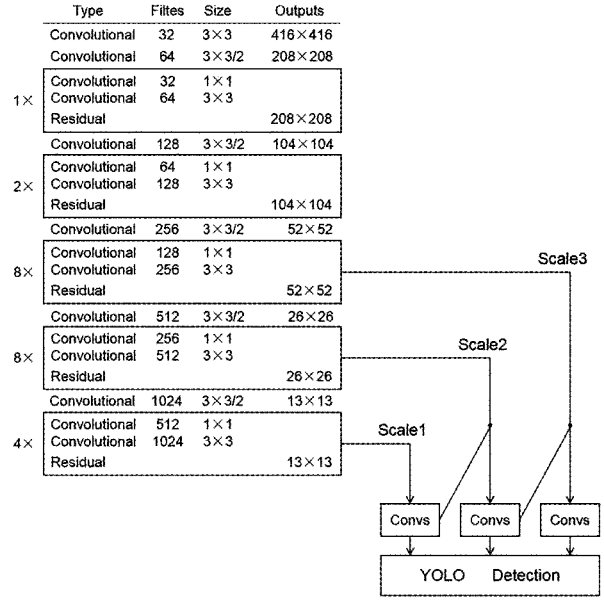
基于深度学习的手势识别方法虽然能够取得较高的识别精度，但随着网络层数的不断加深对神经网络模型在嵌入式设备上存储的硬件成本、计算量和训练运行的难度上带来了巨大挑战。因此，本文提出一种具有轻量级卷积神经网络的手势识别方法，在保持高精度的同时，降低模型大小，使模型更利于在资源有限的移动端或嵌入式设备上的部署。

1　YOLOv3 算法

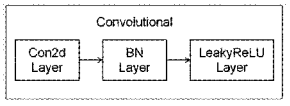
YOLO 是最早的 One Stage 系列模型之一，它构建了一个端到端回归的 CNN 模型。YOLOv3 算法只需要在输入端输入图像数据即可在输出端得到一个预测结果，预测结果为边界框的位置信息、置信度以及所属类别。

1.1　基本原理

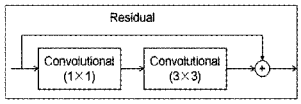
YOLOv3[16] 提出了一种新的特征提取网络结构 Darknet53，该结构由一系列的 1×1 和 3×3 的卷积层组成，每个卷积层后都会跟 1 个 BN 层，对卷积层的输出进行归一化，再输入到激活层，激活函数采用 LeakyReLU，即构成 Darknet53 的最小组件 Convolutional 如图 1(b)所示。Darknet53 共有 5 个步长 2 的卷积层，使用步长为 2 的卷积来代替最大池化进行下采样，降低了池化带来的梯度负面效果，特征图经过 5 次下采样后，416×416 的输入分辨率最终输出 13×13 的特征图。此外，Darknet53 借鉴 ResNet[17] 的思路在网络中引入了大量的残差连接如图 1(c)所示，在增加网络的深度的同时有效抑制了反向传播时梯度爆炸或消失的情况。



(a) YOLOv3网络结构



(b) Convolutional单元



(c) 残差单元

图 1　YOLOv3 网络结构

在检测网络上 YOLOv3 加入了特征金字塔网络(FPN)[18]，将 Darknet53 的 5 次下采样中的后 3 次分别传输到检测网络进行 3 次不同尺度的检测。其中：特征图 Scale3 尺寸最大，负责预测小尺寸物体；Scale1 尺寸最小，负责预测大尺寸物体。YOLOv3 其具体网络结构如图 1 所示。

1.2 损失函数

YOLOv3 的损失函数有 3 部分组成：由中心点和宽高坐标部分带来的边界框误差 $Loss_{box}$、置信度误差 $Loss_{obj}$ 和分类误差 $Loss_{cls}$。其中在边界框损失上采用均方误差计算，在置信度损失和分类损失上采用交叉熵误差计算，最后将 3 个损失求和。YOLOv3 其损失函数具体公式如式(1)所示。

$$
\begin{aligned}
Loss &= Loss_{box} + Loss_{obj} + Loss_{cls} = \\
&\lambda_{coord}\sum_{i=0}^{S\times S}\sum_{j=0}^{B} I_{ij}^{obj}\left[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2\right]+ \\
&\lambda_{coord}\sum_{i=0}^{S\times S}\sum_{j=0}^{B} I_{ij}^{obj}(2-w_i\times h_i)\left[(w_i-\hat{w}_i)^2+(h_i-\hat{h}_i)^2\right]- \\
&\lambda_{obj}\sum_{i=0}^{S\times S}\sum_{j=0}^{B} I_{ij}^{obj}\left[\hat{C}_i\log(C_i)+(1-\hat{C}_i)\log(1-C_i)\right]- \\
&\lambda_{noobj}\sum_{i=0}^{S\times S}\sum_{j=0}^{B} I_{ij}^{noobj}\left[\hat{C}_i\log(C_i)+(1-\hat{C}_i)\log(1-C_i)\right]- \\
&\sum_{i=0}^{S\times S}\sum_{j=0}^{B} I_{ij}^{obj}\sum_{c\in classes}\left[\hat{p}_i(c)\log(p_i(c))+(1-\hat{p}_i(c))\log(1-p_i(c))\right]
\end{aligned}
\tag{1}
$$

式中：S 是 3 个预测层网格的大小，有 13、26 和 52 这 3 个值；B 代表每个网格上的 B 个锚框；λ_{coord} 是边界框预测的惩罚系数；λ_{obj} 和 λ_{noobj} 分别表示包含和不包含目标时置信度的惩罚系数；I_{ij}^{obj} 和 I_{ij}^{noobj} 分别表示第 j 个候选边界框所在的第 i 个网格负责和不负责检测该目标；(x_i,y_i,w_i,h_i) 和 $(\hat{x}_i,\hat{y}_i,\hat{w}_i,\hat{h}_i)$ 分别表示预测边界框和真实边界框的中心点横坐标、纵坐标，边界框的宽度和高度；C_i 和 $\hat{C}_i$ 分别表示预测置信度和真实置信度；$classes$ 表示目标类别的数量；$p_i(c)$ 和 $\hat{p}_i(c)$ 分别表示网格中的目标属于某个类别 c 的预测概率和真实概率。

2 算法改进

2.1 主干网络的改进

YOLOv3 主干网络中使用残差结构增大神经网络的深度实现特征提取，它们会占用很大部分的硬件资源，目前主流的 MobileNet 以及 ShuffleNet 等轻量级模型，采用深度可分离卷积或者分组卷积和通道混洗操作，通过减小卷积核尺寸的方式，构建有效的 CNN 网络，但是其中 1×1 卷积层仍然会大量的消耗内存，增加 FLOPs。GhostNet[19]是华为诺亚方舟实验室提出的一种新型的轻量级网络架构，在保持同样精度的情况下，该模型的计算量和参数量都少于当下最先进的模型。该网络在已有的卷积神经网络基础上引入全新的 Ghost 模块，通过廉价操作生成更多的特征图。Ghost 模块分为两个部分，先通过有限的普通卷积得到一部分特征图，再将这里得到的特征图通过线性操作生成更多特征图，最后将两组特征图在指定维度拼接起来，通过“少量传统卷积计算”+“轻量的冗余特征生成器”的方式，既能保证特征冗余性从而保证精度，又能减少网络的整体计算量。其原理如图 2 所示。

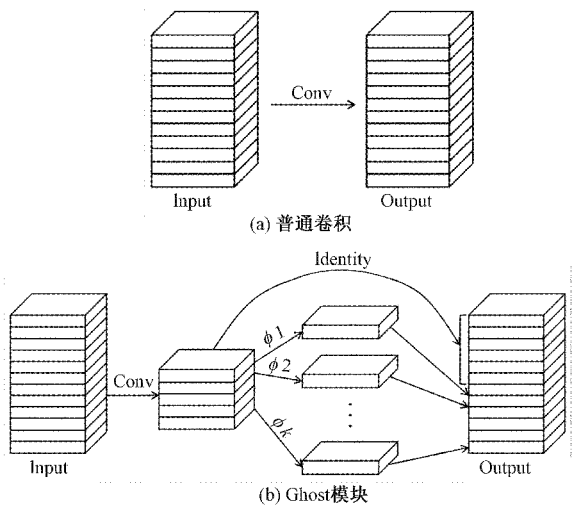


图 2　普通卷积与 Ghost 模块

此外，由于 ReLU 激活函数在输入为负时存在神经元死亡、梯度消失等问题，因此，将 Ghost 模块中 ReLU 函数替换为 SiLU 激活函数。SiLU 激活函数曲线更为平滑，对非线性输出的拟合能力更为优秀。除此之外，SiLU 激活函数不存在斜率接近于 0 的区域，因此不存在梯度消失的现象，而其在负半轴依旧会有输出，因此也不存在 ReLU 的神经元死亡问题。

以 Ghost 模块为基础的 Ghost Bottleneck 如图 3 所示。每个 Ghost Bottleneck 中包含两个 Ghost 模块，第 1 个 Ghost 模块作为扩展层来增加通道维度，第 2 个 Ghost 模块用于压缩特征通道维度以和短连接路径的通道维度匹配。此外，Ghost Bottleneck 有 Stride=1 和 Stride=2 两个版本，分别如图 3(a)、(b)所示。

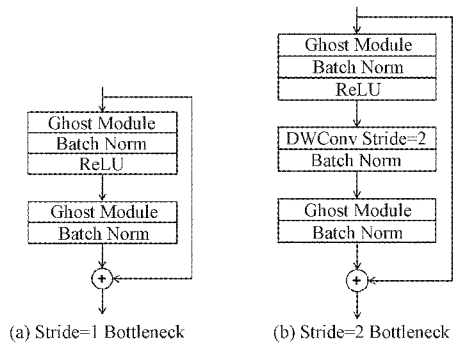


图 3　Ghost Bottleneck

针对基于深度学习的手势识别模型参数量大、训练速度缓慢且对设备要求高，增加了成本的问题，本文提出了一种基于 YOLOv3 的轻量级卷积神经网络结构（以下简称 YOLOv3-GL）。YOLOv3 的主干特征提取网络中使用了多次堆叠的残差单元，这些残差单元是导致 YOLOv3 网络参数量、计算量较大的主要原因，为了减少的参数量和计算量，在不降低精度的前提下，加快 YOLOv3 的前向运算速度，本文对 YOLOv3 的主干特征提取网络进行优化。首先在 YOLOv3 的 5 次下采样中减少残差单元的使用次数，从而减小网络的深度，其次减少每层的通道数也就是特征图。为了进一步减轻对硬件资源的需求，在 YOLOv3 的基础上借鉴了 Ghost 模块的思想，通过使用 Ghost 模块和 Ghost Bottleneck 结构重新设计主干特征提取网络，减少了网络参数量和计算量。

2.2　特征融合的改进

在进行了主要特征的提取后，采用加权双向特征金字塔网络（BiFPN）对主干网络中提取出的特征进行加强，更好地检测不同大小的目标，提升算法检测能力。

YOLOv3 模型中特征融合网络采用的特征金字塔网络（FPN）来检测不同大小的物体，将大尺度特征图与小尺度特征图下采样结果进行自上到下合并，同时获得多个尺度的回归和预测结果。FPN 结构如图 4(a)所示，自上而下对不同级别特征（$P_3 \sim P_7$）进行相加融合。然而，在实际检测过程中不同尺度的特征信息对于最终预测结果的贡献具有差异性，这种差异性通常会对输出特征产生不同的影响。谷歌在 2019 年推出的 EfficientDet 目标检测中提出了一种新型的多尺度特征层融合的结构 BiFPN[20]，在传统的 FPN 中加入了跳跃连接，还将 FPN 中只进行自上到下特征融合替换为自上到下和自下到上的特征融合，使网络在不增加计算额外参数的同时能够融合更多相同尺度和不同尺寸的特征，此外，对于只有一个输入的结点将减少此结点对特征网络的贡献。BiFPN 有效地融合不同尺度的特征并增加了同一尺度特征的信息融合[21]，其结构设计如图 4(b)所示。

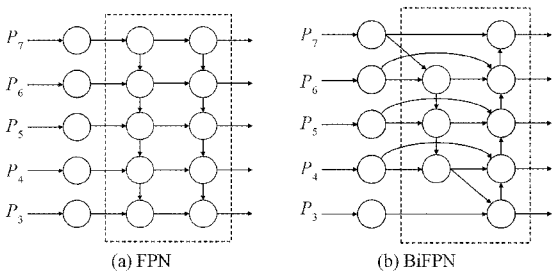


图 4　特征融合结构

输入特征图在加权双向特征金字塔结构中分别经历了自上向下和自下向上两组卷积层处理，其中自上向下的处理按照式(2)进行：

$$\begin{aligned} P_7^{out} &= Conv(P_7^{in}) \\ P_6^{out} &= Conv(P_6^{in} + Resize(P_7^{out})) \\ &\cdots \\ P_3^{out} &= Conv(P_3^{in} + Resize(P_4^{out})) \end{aligned} \tag{2}$$

自下向上的处理按照式(3)、(4)进行：

$$P_6^{td} = Conv\left(\frac{\omega_1 P_6^{in} + \omega_2 Resize(P_7^{in})}{\omega_1 + \omega_2 + \varepsilon}\right) \tag{3}$$

$$P_6^{out} = Conv\left(\frac{\omega'_1 P_6^{in} + \omega'_2 P_6^{td} + \omega'_3 Resize(P_5^{out})}{\omega'_1 + \omega'_2 + \omega'_3 + \varepsilon}\right) \tag{4}$$

其中，P_6^{td} 是第 6 层自上向下路径上的中间特征，而 P_6^{out} 是第 6 层自下向上路径上的输出特征，P_6^{in} 是第 6 层对应的输入特征，所有其他特征均以类似方式构造；ω 为各级特征对应的加权系数；$Conv(\cdot)$ 为特征处理操作；$Resize(\cdot)$ 为分辨率的上下采样操作；ε 为稳定系数。

2.3　损失函数的优化

YOLOv3 模型中使用交叉熵作为置信度和分类的损失函数、均方误差作为边界框回归的损失函数。然而，均方误差损失函数没有考虑到目标框信息间的相关性，容易导致定位不准确，并且在训练刚开始阶段会出现梯度消失现象，导致模型训练速率十分缓慢。为使模型在训练过程中具有更好的收敛性、实现更高效的检测，采用 CIoU[22] 作为边界框回归的损失函数，充分考虑了目标与检测框之间的中心点距离，使预测框更接近真实框，加快模型收敛速度，能够使得网络训练更快的得到最优的预测参数模型。

CIoU 损失函数计算过程如式(5)～(8)所示。

$$L_{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{5}$$

$$L_{CIoU} = 1 - L_{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\nu \tag{6}$$

$$\nu = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h}\right)^2 \tag{7}$$

$$\alpha = \frac{\nu}{(1 - L_{IoU}) + \nu} \tag{8}$$

式中：L_{IoU} 为预测框和真实框之间的交并比损 L_{CIoU} 为 CIoU 损失函数；$\rho^2(b, b^{gt})$ 为预测框中心点 b 和真实框中心点 b^{gt} 之间的欧氏距离；c 为同时包含预测框和真实框的闭合矩形对角线距离；α 为平衡尺度权重参数；ν 为衡量预测框和真实框的长宽比相似性参数；ω^{gt}、h^{gt} 分别为真实框的宽和高；ω、h 分别为预测框的宽和高。

2.4　YOLOv3-GL 网络

本文提出的改进方案从原始 YOLOv3 模型的输入端、主干网络、颈部和检测头 4 个方面进行优化：输入端引入 Mosaic 数据增强处理；主干网络中减少残差单元的使用次数和每层的通道数，同时，使用 Ghost 模块和 Ghost Bottleneck 结构重新设计特征提取网络，并将 Ghost 模块中的 ReLU 激活函数替换为 SiLU 激活函数；颈部采用加

权双向特征金字塔结构进行多尺度特征融合；检测头部分将原始边界框回归均方误差（MSE）损失函数替换为 CIoU 损失。改进后的 YOLOv3 轻量化网络结构（YOLOv3-GL）整体如图 5 所示。

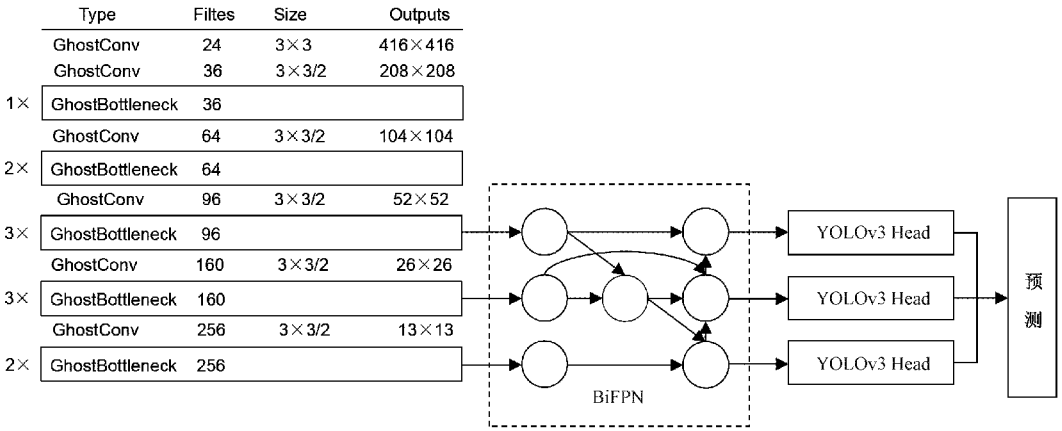


图 5　YOLOv3-GL 网络结构

经过分析计算，改进后的 YOLOv3-GL 模型总计 466 层，4 691 349 个参数量，计算量为 13.0GFLOPs，与改进前 YOLOv3 模型相比，优化后的模型参数量减少为原来的 7.62%，计算量减少为原来的 8.38%，模型大小仅为 17.9 MB，实现了对原模型大程度的压缩。

3　实验与结果分析

3.1　实验数据

实验使用的数据集是公开的 NUS-II 手势数据集（Pisharady 等，2013）。这是一个 10 类手部姿势数据集，这些手部姿势是在新加坡国立大学和它的周围拍摄的，拍摄背景是复杂的自然背景。该手势数据集是由 40 名不同种族、年龄在 22～56 岁之间、肤色存在巨大差异的采集者在不同复杂的背景下完成的，每一名采集者分别展示字母“a”到“j”总共 10 种手势，每个字母手势拍摄 5 次，共包含 2 000 张 RGB 图片。

同时，数据集中还包括带有人类干扰背景的 750 张 RBG 图片，背景包含行人和人脸等，数据集共计 2 750 张。将手势数据集随机取 2 475 张图像作为训练样本，275 张图像作为测试样本，并对 NUS-II 所有图片中的手势利用标记软件 LabelImg 进行人工标注，完成数据集准备工作，最终得到格式为 xml 的标签格式，将制作好的数据集保存为 PASCAL VOC 格式，以便于网络的训练与测试。实验数据集部分图片如图 6 所示。

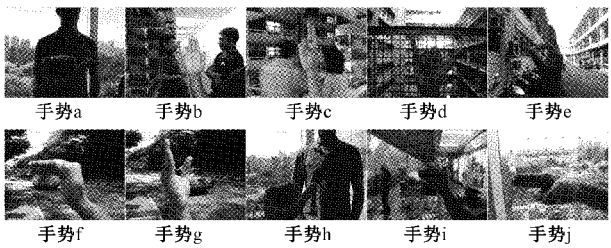


图 6　部分手势数据集

3.2　数据增强

随机改变训练样本可以降低模型对物体出现位置的依赖，提高模型的泛化能力，因此算法在训练过程中对训练数据进行 Mosaic 数据增强训练技巧。

Mosaic 数据增强的主要过程是：1）随机读取每个批次中的 4 张图像；2）分别对 4 张图像进行翻转、缩放、平移等操作，并按照左上、右上、左下和右下的位置放置；3）裁剪拼接形成新的图片用于训练，并且只保留裁剪后还完整的标注框。

输入端的 Mosaic 数据增强大大丰富了检测物体的背景，增加了数据集的多样性，随机翻转、缩放、平移等操作增加了不同形态的目标，增强了网络的鲁棒性。此外，将 4 张图片拼接的方式变相地提高了训练过程中每批次输入图片的数量，使得训练的超参数 Batch Size 不需要设置地很大，间接地降低了对显存的要求，单个 GPU 就可以达到很好的效果。

3.3　实验环境

实验机器操作系统版本为 Windows10，CPU 型号为 Intel Xeon CPU E5-2678 v3 @ 2.50 GHz，GPU 型号为 NVIDIA GeForce RTX 2080 Ti，显存大小为 11 G，内存大小为 62 G。采用 Python 作为编程语言，基于 Pytorch1.8 深度学习框架，并使用 cuda11.1 和 cudnn8.0.5 对 GPU 进行加速。

3.4　评价指标

采用的算法评价指标为准确率（Precision）、召回率（Recall）、mAP@0.5、mAP@0.5∶0.95、网络推理时间、运算浮点数（FLOPs）、参数量（Parameters）和模型大小（model size）。mAP 计算公式如式（9）～（12）所示。

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$AP = \int_0^1 P(R)\mathrm{d}t \tag{11}$$

$$mAP = \frac{1}{N}\sum_{n=0}^{N} AP_n \tag{12}$$

式中：P 表示准确率，R 表示召回率，TP 代表的是预测为正的样本，FN 代表预测为负的样本，FP 代表预测为正的样本；AP 被定义为 PR 曲线下的面积，用来衡量数据集中一类的平均分类精确率；mAP@0.5 是指 IoU 设为 0.5 时所有类别的平均 AP，mAP@0.5∶0.95 是指在不同 IoU 阈值上的平均 mAP，IoU 取值从 0.5～0.95，步长为 0.05；FLOPs 用来衡量算法模型的复杂度；Parameters 表示网络中的参数数量的多少；模型大小是指最终训练结束得到保存的模型大小。

3.5　网络训练

将数据导入算法模型进行训练，在模型训练的过程中使用 SGD 优化方法。模型训练的超参数设置：初始学习率设置为 0.005，动量设置为 0.9，权重衰减系数设置为 0.000 5，在实验的训练和测试中所设置的图片大小设置为 416×416，batch-size 大小设置为 16，epoch 大小设置为 100。

3.6　结果与分析

YOLOv3-GL 模型在训练时损失值函数 Loss 的变化趋势如图 7 所示。

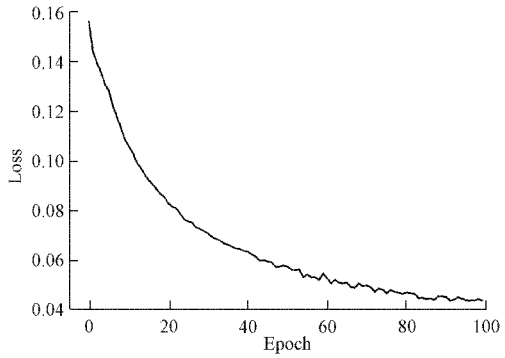


图 7　Loss 变化曲线

通过图 7 结果可以看出，损失函数在训练的初期下降迅速，表明模型正在快速拟合，模型的学习效率较高，当 Epoch 为 40 时，模型损失值减小速度开始变缓，当 Epoch 为 80 时，Loss 曲线逐渐平缓，损失值在 0.04 附近波动，模型达到稳定状态。

YOLOv3-GL 模型在训练时平均精确度的变化趋势如图 8 所示。

通过图 8 结果可以看出，在模型训练 30 个 Epoch 后，10 类手势在测试集上的平均精确度已达到 80%附近，并随着 Epoch 增加而稳步上升，40 个 Epoch 后趋向平稳。

为了验证 YOLOv3-GL 中所用的各个模块对 YOLOv3 的优化作用，进行了消融实验。首先使用 Ghost 模块设计主干特征提取网络，之后分别在该模型的基础上分别单独添加 BiFPN 结构、CIoU 损失函数和 Mosaic 数据增强，表 1 为对 YOLOv3-GL 模型所做的消融实验结果，分别比较了各种结构组合下的算法平均精确度和模型大小。

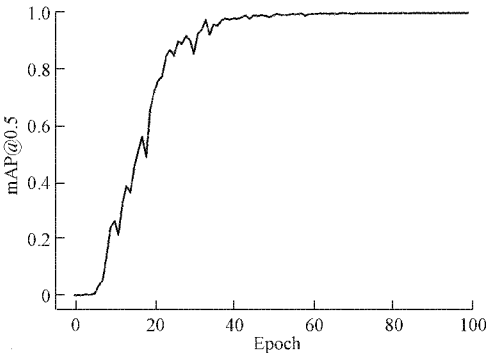


图 8　平均精确度变化曲线

表 1　消融实验

实验	主干网络改进	BiFPN	CIoU	Mosaic	mAP/%	模型大小/MB
YOLOv3	×	×	×	×	98.9	235
改进模型 1	√	×	×	×	96.9	57.3
改进模型 2	√	√	×	×	97.3	17.9
改进模型 3	√	√	√	×	98.3	17.9
YOLOv3-GL	√	√	√	√	99.5	17.9

可以看出 YOLOv3 模型的权重文件大小为 235 MB，mAP 为 98.9%；改进模型 1 对 YOLOv3 网络重新设计轻量化主干特征提取网络，其权重文件大小降低了 177.7 MB，模型大小为 57.3 MB，但 mAP 相较于 YOLOv3 模型降低了 2%；改进模型 2 引入 BiFPN 改进特征融合网络，将权重文件大小降低至 17.9 MB，mAP 相较于改进模型 1 提高了 0.4%，表明利用不同特征的权重信息有利于提高模型检测精度；改进模型 3 采用 CIoU 作为边界框回归损失函数，相较于改进模型 2，mAP 提高了 1%，表明 CIoU 作为边界框回归损失函数可改善模型检测精度；改进后的 YOLOv3-GL 模型在输入端引入 Mosaic 数据增强，权重文件大小仅为 17.9 MB，mAP 为 99.5%，相较于 YOLOv3 模型权重文件大小降低了 92.4%，mAP 提高了 0.6%，表明改进后的模型能够大幅减少原 YOLOv3 模型大小并提高检测精度。

为进一步验证改进后的 YOLOv3-GL 算法在轻量化网络方面的优势，将改进后的网络结构与其他算法进行对比，实验结果如表 2 所示。

实验结果表明，相较于 YOLOv3，使用 Ghost 卷积设计的轻量级主干特征提取网络，改进后的算法具有更少的参数量和计算量，同时，由于引入 BiFPN 结构和 CIoU 损

表 2　YOLOv3-GL 与其他算法对比

实验	Params	GFLOPs	模型大小/MB	推理时间/ms	mAP/%
YOLOv3	6.16×10^7	155.2	235	16.06	98.9
YOLOv3-Tiny	8.69×10^6	13	33	4	85.6
YOLOv5s	7.05×10^6	15.9	27	9.07	99.3
SSD	2.42×10^7	31.4	92.1	20.73	99.1
YOLOv3-GL	4.69×10^6	13	17.9	10.86	99.5

失函数，并取得了更高的准确率和更快的网络推理时间。

此外，YOLOv3-Tiny 算法是 YOLOv3 算法的轻量级版本，YOLOv3-Tiny 算法在 YOLOv3 的基础上去掉了大量的卷积层和残差结构，在多尺度检测上只保留了 13×13 和 26×26 两个预测分支，采用上采样的方式将 13×13 的特征层和 26×26 的特征层进行融合，最后在两个特征图上分别作出预测。因此，虽然网络推理时间更短但是检测准确率大幅降低，在 NUS-II 手势数据集上仅为 85.6%，不能满足手势识别检测的需求。

SSD 算法使用 VGG16 作为主干特征提取网络，并使用 6 个尺度融合的检测方式，但 SSD 的参数量和模型大小都比较大且网络推理时间较长，性能远不如本文算法。

YOLOv5 一共有 4 个版本，分别为 YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x，其中 YOLOv5s 是深度最小，特征图宽度最小的网络，模型大小仅为 27 MB，但本文算法比 YOLOv5s 参数量、计算量和模型大小更小，且检测精度更高。

以上实验表明，改进后的 YOLOv3-GL 算法，对手势识别检测的平均准确率达到了 99.5%，比 YOLOv3 高 0.6%，比 YOLOv3-Tiny 高 13.9%，比 YOLOv5s 高 0.2%，比 SSD 高 0.4%，模型大小仅为 17.9 MB，比 YOLOv3 低 217.1 MB，比 YOLOv3-Tiny 低 15.1 MB，比 YOLOv5s 低 9.1 MB，比 SSD 低 74.2 MB，具有较高的检测精度，同时使网络轻量化。测试集的部分检测结果如图 9 所示。

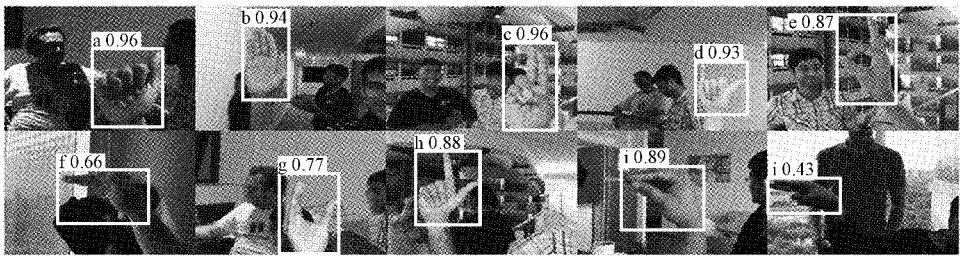


图 9　基于 YOLOv3-GL 的手势识别检测结果

4　结　　论

针对基于深度学习的手势识别模型参数量大、训练速度缓慢且对设备要求高，增加了成本的问题，针本文提出了一种基于轻量级卷积神经网络的手势识别检测算法 YOLOv3-GL。利用 Ghost 卷积模块设计轻量级主干特征提取网络，解决了原来网络计算量和参数量大的问题；通过引入加权双向特征金字塔网络改进特征融合网络，提升网络检测精度；最后使用 CIoU 损失函数作为边界框回归损失函数并加入 Mosaic 数据增强技术，加快模型收敛速度提升网络的鲁棒性。实验结果表明，提出的方法有效地对 YOLOv3 模型进行了压缩并加快推理速度，大大减少了模型的参数量和计算量，模型大小仅为 17.9 MB，精度能够达到 99.5%，相比于原 YOLOv3 模型参数量减少 92.38%，模型计算量减小 91.62%，平均精确度提高了 0.6%，在 GPU 上的推理时间为 10.86 ms。

因此改进后的算法有效地实现了准确快速的手势识别检测，改善了训练速度缓慢和内存占用大的缺点，使模型更有利于搭载在嵌入式设备上。

参考文献

[1]　冯志全，蒋彦. 手势识别研究综述[J]. 济南大学学报，2013，27(4)：336-341.

[2]　卢梦圆，宫巍，马力. 基于多特征融合的手势识别研究[J]. 计算机与数字工程，2020，48(9)：2157-2161.

[3]　王剑波，朱欣娟，吴晓军. 融合静态手势特征和手部运动轨迹特征的手势交互方法[J]. 国外电子测量技术，2021，40(7)：14-18.

[4]　程淑红，程彦龙，杨镇豪. 基于手势多特征融合及优化 Multiclass-SVC 的手势识别[J]. 仪器仪表学报. 2020，41(6)：225-232.

[5]　舒子超，曹晓松，谢代梁，等. 基于三维视觉特征的数字手势语义识别新方法研究[J]. 电子测量与仪器学报，2021，35(6)：124-130.

[6]　REDMON J，DIVVALA S，GIRSHICK R，et al. You only look once：Unified，real-time object detection[C]. IEEE，CVPR2016 Conference on Computer Vision and Pattern Recognition，Washington DC：IEEE Computer Society Press，2016：779-788.

[7]　REDMON J，FARHADI A. YOLO9000：Better，faster，stronger[C]. IEEE. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，Washington DC：IEEE Computer Society Press，2017：6517-6525.

[8] MOLCHANOV P, MALLYA A, TYREE S, et al. Importance estimation for neural network pruning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 11264-11272.

[9] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]. Proceedings of the European Conference on Computer Vision, Berlin, Heidelberg: Springer, 2016: 21-37.

[10] GIRSHICK R, DONAHUE J, DARRELL T, et al. Region-based convolutional networks for accurate object detection and segmentation [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[11] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[12] 谢淋东，仲志丹，乔栋豪，等. 多尺度卷积特征融合的 SSD 手势识别算法[J]. 计算机技术与发展，2021，31(3)：100-105.

[13] 张雷乐，田军委，刘雪松，等. 一种改进的残差网络手势识别方法[J]. 西安工业大学学报，2021，41(2)：206-212.

[14] 袁帅，韩曼菲，张莉莉，等. 基于改进 YOLOV3 与贝叶斯分类器的手势识别方法研究[J]. 小型微型计算机系统，2021，42(7)：1464-1469.

[15] 郭紫嫣，韩慧妍，何黎刚，等. 基于改进的 YOLOV4 的手势识别算法及其应用[J]. 中北大学学报(自然科学版)，2021，42(3)：223-231.

[16] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. ArXiv Eprints, 2018, ArXiv:1804.02767.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016: 770-778.

[18] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017: 936-944.

[19] 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.

[20] TAN M, PANG R, LE Q V. EfficientDet: Scalable and efficient object detection [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020: 10781-10790.

[21] HU L, JIANG Y, LI J, et al. Change detection method for high-resolution remote sensing image based on multi-scale and sparse convolution[J]. Journal of Chinese Computer Systems, 2020, 41(11): 2365-2370.

[22] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding bix regression [C]. AAAI Conference on Artificial Intelligence, New York, 2020: 12993-13000.

作者简介

牛雅睿，硕士研究生，主要研究方向为计算机视觉。
E-mail:nyr98446@163.com

武一(通信作者)，博士，教授，主要研究方向为智能控制系统研究与应用。
E-mail:wuyihbgydx@163.com

孙昆，硕士研究生，主要研究方向为计算机视觉。
E-mail:skypai1011@163.com

卢昊，硕士研究生，主要研究方向为无人系统智能感知。
E-mail:lh0102251@outlook.com

赵普，硕士研究生，主要研究方向为计算机视觉。
E-mail:zp_dling@163.com