基于 MEF-YOLO 的轻量手势识别算法

朱雪燕[1] 王招娣[1] 黄明茹[1] 郭梦珏[1]
ZHU Xueyan WANG Zhaodi HUANG Mingru GUO Mengjue

摘 要 针对非接触式的人机交互中手势识别精度低速度慢的问题，提出一种轻量化 MEF-YOLO（MobileNetV3-ECA-FReLU YOLO）算法。将 YOLOv5s 的主干网络 CSPDarknet53 替换成轻量化的 Mobielnetv3，在主干网络的最后一个卷积层之后融入 ECA 注意力机制，规避因参数减少而导致的特征信息丢失问题，同时使模型更好地融合不同通道间的信息，接着在输出层添加 FReLU 激活函数，增加模型的非线性，使特征的表达能力增强。在自制数据集上验证了 MEF-YOLO 算法的可行性，并与 YOLOv5 算法进行了对比。结果表明，轻量化 MEF-YOLO 算法的模型大小减小了 78.4%，检测速度提升了 61 帧 /s，同时平均识别精度较 YOLOv5 算法提升了 3.6%。

关键词 手势识别；MobileNetV3；注意力机制；FReLU 激活函数

doi：10.3969/j.issn.1672-9528.2024.05.003

0 引言

手势识别作为一种人与人之间的非接触式交互方式，具有直接性、自然性、高效性，提高了人机信息的交互效率。传统的手势识别方法通过提取手掌特征来进行目标检测，但检测效果易受光线和环境的干扰，存在对手势变化适应性差、处理速度慢的缺点。随着深度学习在计算机领域的发展，手势识别作为直观方便的人机交互方式引起各界的广泛关注。Mujahid 等人 [1] 提出了一种基于 YOLOv3 和 DarkNet-53 深度学习模型的轻量级手势识别模型；Bochkovskiy 等人 [2] 在 YOLOv4 的基础上加入了加权残差连接（WRC）和跨阶段局部连接（CSP）等功能，提高了目标实时监测速度。研究表明，YOLO 算法 [3] 在手势识别领域具有广泛的应用。YOLOv5 虽在专业领域内得到一定的发展，但仍存在对设备硬件要求较高、实时性和准确率之间不平衡的问题。因此，为解决该问题，本文对 YOLOv5 在模型运行速度和检测精度方面进行了一系列改进，提出了 MEF-YOLO 算法。经验证，该算法在目标检测中是可行的，能在保持高速检测的同时提高检测精度，且其所占用的内存小，为手势识别技术的发展做出了积极的贡献。

1. 洛阳师范学院 河南洛阳 471934

[基金项目] 河南省科技厅科技攻关项目 (222102210301)；国家级大学生创新创业训练计划项目（202310482024）；国家级大学生创新创业训练计划项目（202310482010）；洛阳师范学院高等教育教学改革研究与实践项目（2023XJGJ025）；2023 年河南省产教融合重点项目（教办高〔2024〕13 号文件，项目序号 30）

1 YOLOv5s 改进算法

文中在 YOLOv5[3-4] 基础上进行了一系列改进。首先，引入了 MobileNetV3 作为网络的主干部分，以提高模型的轻量化和运算效率。MobileNetV3[5-6] 采用了非线性瓶颈模块，能够在保持准确性的同时大幅度减少计算量。其次，引入了注意力机制 ECA（efficient channel attention），以增强网络对不同空间尺度和通道之间的特征关联性的学习能力。然后，ECA 可以自适应地调整通道间的相关性，提升了网络在手势图像中捕捉关键信息的能力，进而提高了模型的检测精度和鲁棒性。最后，采用了一种新的激活函数 FReLU（fast rectified linear unit），该函数在快速计算和保持线性单调性的基础上，能更好地表达手势图像中的非线性特征，进一步提升模型的性能和准确性。

1.1 MobileNetV3 轻量型主干网络

MobileNetV3 采用了一种倒残差结构，即先将输入通过一个 1*1 卷积降维，然后通过一系列的 3*3 或 5*5 深度可分离卷积层进行特征提取，最后通过 1*1 卷积进行通道扩展，将特征图的通道数再次扩展，以便进行后续的特征处理和分类。在保证较高的准确率的同时，有效地减少了计算量和参数量，提高了检测速度。

本文采用 MobileNetV3 作为 YOLOv5 的主干网络，将输入的图像经过一系列卷积层和池化层提取出高层次的语义特征，然后传递给 YOLO head 网络，用于预测目标的边界框和类别信息。MobileNetV3 引入了非线性瓶颈模块和 SE 模块两种新的网络模块。非线性瓶颈模块采用了一种非线性激活函

数（如 h-swish 函数）来替代传统的 ReLU 函数[7-8]，提高了模型的非线性表达能力。

ReLU6 激活函数如式（1）所示，相当于加了个最大值 6 进行限制。

$$y = \mathrm{Re}LU6(x) = \min(\max(x,0),6) \tag{1}$$

hardswish 激活函数如式（2）所示，相当于分成 3 段进行限制。采用 *hardswish*，计算速度相对较快，对量化过程友好。

$$Hardwish(x) = \begin{cases} 0 & if\ x \le -3 \\ x & if\ x \ge +3 \\ x \cdot (x+3)/6 & otherwize \end{cases} \tag{2}$$

SE 模块对特征图进行全局平均池化以及先降维后升维两个全连接层得到输出向量。特征图经过两个全连接层之后，比较重要的特征图对应的向量元素的值较大。将得到的权重和对应特征图中的所有元素相乘，得到新的输出特征图。通过一系列的深度可分离卷积层和其他模块，对图像进行语义信息提取和特征提取，送入 YOLOv5 的检测头进行目标检测，轻量化处理后模型如图 1 所示。融入 MobileNetV3 后的 YOLOv5 在目标检测的准确性和鲁棒性方面提升很大。

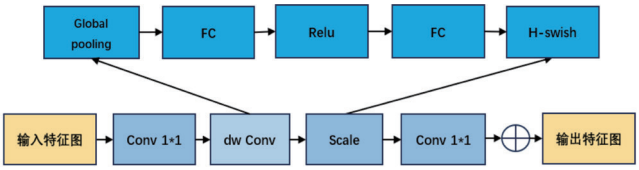


图 1 融合 MobileNetV3 后的 YOLOv5 结构图

1.2 增加注意力机制 ECA

YOLOv5 结合 MobileNetV3 后作为一种轻量级目标检测算法，其网络结构相对简单，参数量变少，无法充分表达图像中复杂的语义信息，特征网络提取到的特征信息也变少。同时，YOLOv5 通过不同层次的特征融合来实现多尺度目标检测，但传统的融合方法可能无法充分利用不同尺度特征之间的关联性，从而导致检测精度降低。因此，本文通过引入一种有效的注意力机制 ECA 来提高模型对重要特征的关注程度，增强特征的表达能力，ECA 模块自适应地调整特征图中每个通道的权重，使得模型能够更好地融合不同尺度的特征，有效捕获了跨通道交互的信息，提高了目标检测的精度和鲁棒性。

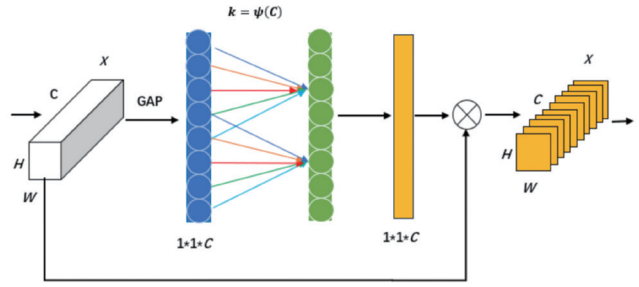


图 2 ECA 结构图

如图 2 所示，ECA 注意力机制首先通过全局平均池化层对输入的 C 个特征图进行操作，将特征图从 $[H,W,C]$ 的矩阵变成 $[1,1,C]$ 的向量，得到每个通道的全局信息；接着将维度为 $1\times1\times C$ 的特征向量进行 1*1 卷积，并经过 Sigmoid 非线性激活函数，得到每个通道的权值；最后，将注意力权重向量与原始特征图逐通道相乘，得到经过注意力加权的特征图，减少对无关信息的关注。

ECANet 是对 SENet（squeeze-and-excitation network）的改进[9-10]。相比于 SENet，ECA 注意力机制模块直接在全局平均池化层之后使用卷积层，去除了全连接层，避免了降维对注意力机制的副作用，并有效捕获了跨通道交互。ECA 能够根据卷积核的大小通过一个函数来自适应变化，调整特征的权重，从而更好地捕捉图像中的关键信息，提高了计算效率。

ECA 的卷积核的自适应函数为：

$$k = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right| \tag{3}$$

式中：C 是通道维数。

本文通过在卷积层后面插入 ECA 模块，使模型自适应地调整每个通道的重要性，以减少模型对于不重要特征的过度依赖，进而增强每个卷积层的通道注意力和感受野，帮助模型更好地关注重要的特征通道，扩大每个位置的感受野，同时提升特征的表达能力，进一步改善 YOLOv5 的目标检测性能。融合 ECA 后的 YOLOv5 结构图如图 3 所示。

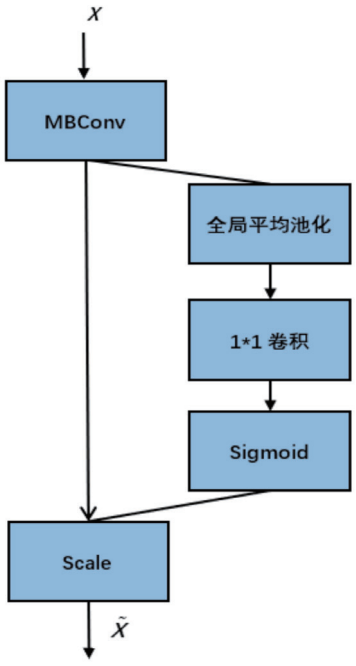


图 3 融合 ECA 后的 YOLOv5 结构图

1.3 采用 FReLU 激活函数

YOLOv5 所采用的 ReLU 激活函数，存在信息丢失和模型性能下降问题。本文选用在 ReLU 基础上改进的 FReLU 激活函数，在负数区域引入可通过学习来确定的可调参数，使函数输出在负数部分有非零值，使得函数在梯度传播时更加灵活，既解决了梯度消失现象，也避免了神经失活现象，进而能提高模型的表达能力。

FReLU 激活函数将输入展平成一个一维向量，对每个元素应用 ReLU 激活函数，再将输出重新恢复成原来的形状，实现了在不增加参数量的情况下提高模型的表示能力和鲁棒性。FReLU 通过对局部卷积后的输出与原始数据进行一个 max 的比对，将 ReLU 与 PReLU 扩展成 2D 的激活函数，即将原先 ReLU 的 $x<0$ 部分换成了 2D 的漏斗条件，增强激活空间的灵敏度，以改善图像视觉。FReLU 在每个通道上学习一个独立的参数，能自适应地学习每个通道的不同非线性形状，提高模型的灵活性。

$$f\left(x_{c,i,j}\right)=\max\left(x_{c,i,j},\mathrm{T}\left(x_{c,i,j}\right)\right)$$
$$\mathrm{T}\left(x_c,i,j\right)=x_{c,i,j}^{\omega}\cdot p_c^{\omega} \tag{4}$$
$$x_{c,i,j}^{\omega}\cdot p_c^{\omega}=\sum_{i-1\le h\le i+1,j-1\le\omega\le j+1}x_{c,h,\omega}\cdot p_{c,h,\omega}$$

式中：$x_{c,i,j}$ 代表需要激活的像素，c,i,j 对应 channel 和 2D 位置。

表达式为：

$$F=\max\left(x,T\left(x\right)\right) \tag{5}$$

式中：$T(x)$ 代表简单高效的空间上下文特征提取器。

2 实验结果和分析

2.1 实验环境和数据集

本文所有实验均在 Windows10 操作系统下执行，采用的处理器为 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30 GHz，16 GB 运行内存，显卡为 NVIDA GeForce RTX 3050 Laptop GPU。本文基于深度学习框架 PyTorch1.10，搭建的软件环境为 Anaconda3.0、PyCharm2021.3、Python3.8。文中共采集了 26 种不同的手势，尺寸均为 640×640，通过对采集的图片进行翻转、缩放、位移、镜像等操作来增广数据集，又通过中值滤波来进行数据增强。扩充后的数据集，共 7800 张图片，按照 7:2:1 的比例划分为训练集、验证集和测试集。

2.2 实验结果与分析

2.2.1 训练结果与分析

MEF-YOLOv5 算法在测试集中各类别的有效性检测结果如表 1 所示，MEF-YOLOv5 算法的精准率 P、召回率 R 和平均精度 P_{MA} 分别达到 94.8%、91.4%、88.0%，能够有效地实现在复杂场景下的手势识别和分类。

表 1 MEF-YOLOv5 算法各类别的准确率、召回率和平均精度检测结果

类别	P/%	R/%	P_{MA}/%
A	92.4	94.6	85.5
B	98.2	100.0	89.5
C	94.4	96.2	88.8
D	96.1	100.0	89.9
E	100.0	99.5	89.5
F	96.4	95.2	88.5
G	92.4	100.0	88.4
H	92.1	94.9	89.0
I	93.2	92.6	87.2
J	100.0	91.7	85.8
K	99.5	93.3	88.8
L	93.7	100.0	89.5
M	89.7	84.6	83.7
N	93.5	65.3	86.5
O	96.6	85.7	86.7
P	97.7	100.0	89.5
Q	96.6	93.6	89.3
R	98.6	82.7	89.2
S	94.7	89.8	84.9
T	89.5	81.0	86.6
U	95.0	90.2	87.8
V	87.0	91.6	87.8
W	98.0	93.3	89.5
X	93.1	72.2	89.2
Y	91.4	100.0	89.5
Z	95.3	89.3	86.3

用 MEF-YOLOv5 算法，经 200 轮训练的模型损失函数 Loss、精度、mAP 的曲线图如图 4 所示。由图 4 可以得到，Loss 曲线在 30 轮以内呈逐渐下降趋势且下降幅度较大，在 30 ～ 100 轮之间下降幅度稍有平缓，在 100 轮之后损失函数逐渐稳定在 0.01 ～ 0.02 内。精度和 mAP 曲线在 40 轮以内呈逐渐上升趋势且上升幅度较大，在 100 轮以后逐渐趋于稳定且接近于 0.98，说明该模型训练准确率和效率都较好，模型训练成功。

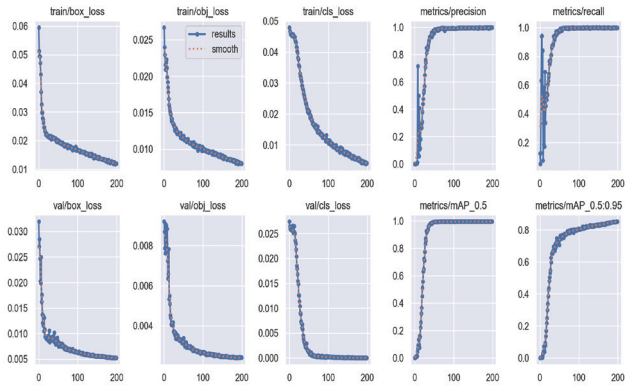
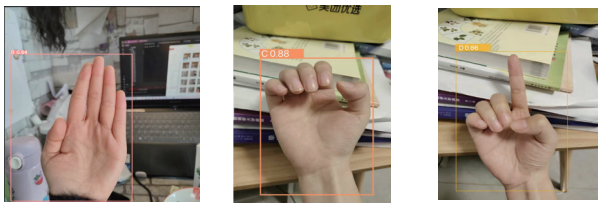


图 4 训练模型的损失函数 Loss 和精度 mAP 的曲线图

在复杂背景下测试，对 26 种手势检测部分识别效果如图 5 所示，手势 B 的识别率为 95%，手势 C 的识别率 90%，手势 D 的识别率为 92%。可以看出，该算法能够准确地检测出在有背景干扰情况下的 A ～ Z 的 26 种手势，能较好地应对干扰，具有较好的鲁棒性。



（a）手势为 B　（b）手势为 C　（c）手势为 D

图 5 复杂环境下的手势检测结果

2.2.2 算法对比与分析

不同模型性能指标对比如表 2 所示，对于 YOLOv5 网络模型，虽然此模型检测精度高，平均精度可达 99.2%，但是模型占用的内存较大，不适合在边缘终端设备上进行部署。将 YOLOv5 的特征提取网络轻量化后，M-YOLOv5 相较 YOLOv5 而言，虽然平均精度下降了 10.6%，但模型占用的内存大大减少，推理速度得到了大幅提高。在 M-YOLOv5 中加入注意力机制模块后，模型的推理速度没有明显变化，模型所占用的内存几乎没有发生变化，但 ME-YOLOv5 的平均精度相较于 M-YOLOv5 提高了 3.8%。同时，本文提出的 MEF-YOLOv5 网络模型采用了 FReLU 激活函数，其平均精度与 M-YOLOv5 相比提升了 7%，并且相较于 YOLOv5，其模型占用的内存是原来的 21.6%，推理速度提高了 61 帧 /s，相比之下效率有了明显提升。YOLOv5 和本文所提出的新算法 MEF-YOLOv5 的平均精度相近，新算法的平均精度损失仅为 3.6%。这也验证了本文所提出的 MEF-YOLOv5 算法的可行性。

表 2 不同算法性能指标对比结果

类别	P/%	R/%	P_{MA}/%	W_{size}/MB	F_{frame}/(帧 ·s^{-1})
YOLOv5	99.1	98.8	99.2	13.7	315
M-YOLOv5	87.3	81.5	88.6	2.97	482
ME-YOLOv5	91.4	83.2	92.4	2.99	482
MEF-YOLOv5	94.8	91.4	95.6	3.47	376

3 结论

本研究提出了一种以 YOLOv5 为基础框架，加入了 MobileNetv3、ECA 和 FReLu 模块的轻量化 MEF-YOLO 手势识别算法。对 26 类手势样本进行测试表明，在多种场景下，该算法能够实现对不同手型和尺寸的手势识别，平均识别精度可达 95.6%。与传统的 YOLOv5 算法相比，轻量化 MEF-YOLO 算法的模型大小减少了 78.4%，检测速度提升了 61 帧 /s，同时平均识别精度较 M-YOLO 算法提升了 7%。本文所提出的 MEF-YOLO 算法为不同场景下的目标跟踪、语义分割、语义理解、物体检测等计算机视觉领域的研究提供了一种新的解决方案。

参考文献：

[1]MUJAHID A, AWAN MJ, YASIN A, et al. Real-time hand gesture recognition based on deep learning YOLOv3 model[J]. Applied sciences,2021,11(9):4164.

[2]BOCHKOVSKIY A, WANG C, MARK L, et al.YOLOv4: optimal speed and accuracy of object detection[EB/OL].(2020-04-23)[2024-01-28].https://arxiv.org/abs/2004.10934.

[3]KÖPÜKLÜ O, GUNDUZ A, KOSE N, et al.Real-time hand gesture detection and classification using convolutional neural networks[C]//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019).Piscataway:IEEE,2019:1-8.

[4]CHEN L, WANG F, DENG H, et al.A survey on hand gesture recognition[C]//2013 International Conference on Computer Sciences and Applications.Wuhan:CSA, 2013:313-316.

[5]MOHAMMED Q A A. 基于深度学习的视觉手势识别方法 [D]. 成都 : 四川大学 ,2021.

[6]WANG Q, WU B, ZHU P, et al.ECA-net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE, 2020:11531-11539.

[7]HOWARD A, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL].(2017-04-17)[2024-01-29].https://arxiv.org/abs/1704.04861.

[8]FU J, LIU J, TIAN H, et al.Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE, 2019:3141-3149.

[9]HAN K, WANG Y, TIAN Q, et al.Ghostnet: more features from cheap operations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2020:1577-1586.

[10]TAN M, PANG R, LE Q.EfficientDet: scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2020: 10778-10787.

【作者简介】

朱雪燕（2002—），女，河南南阳人，本科，研究方向：数字图像处理、深度学习。

王招娣（1988—），女，河南洛阳人，硕士，讲师，研究方向：深度学习、人工智能。

（收稿日期：2024-02-29）