DOI:10.19651/j.cnki.emt.2210441

改进 MobileViT 与 YOLOv4 的轻量化车辆检测网络*

郑玉珩 黄德启
(新疆大学电气工程学院 乌鲁木齐 830017)

摘 要: 基于深度学习的目标检测算法在智能交通的应用中,对于车辆检测存在模型参数量大、计算速度慢和简单网络精准度较低的问题。本文提出了一种高效的轻量化车辆检测模型,该检测模型采用 YOLOv4 网络作为参考模型进行改进。首先,本文采用 CSPMobileViT 网络来替换原始主干网络,然后将 PANet 替换成 BiFPN,并且将 BiFPN 中的 3×3 标准卷积替换成深度可分离卷积,最后,在 BiFPN 之前和 YOLO-Head 之前添加 ECA 模块。在损失函数部分,将边框回归损失 CIoU 改进为 Focal EIoU 来解决难易样本不平衡的问题。实验结果表明改进网络的 mAP 值为 96.77%,检测速度达到每张图片 0.023 4 s,模型大小只有 32.76 MB,参数量为 8 587 541,与原始算法相比 mAP 提升了 1.54%,而模型大小和参数量仅约为原始模型 1/8,并且 FPS 提升了 7.5,改进算法具有更好检测效果。

关键词: YOLOv4;MobileViT;车辆检测;轻量化网络
中图分类号: TP391 **文献标识码:** A **国家标准学科分类代码:** 520.20

Lightweight vehicle detection network based on MobileViT and YOLOv4

Zheng Yuheng Huang Deqi
(School of Electrical Engineering, Xinjiang University, Urumqi 830017, China)

Abstract: In the application of Intelligent Transportation, target detection algorithm based on deep learning has the problems of large number of model parameters, slow calculation speed and low accuracy of simple network for vehicle detection. This paper presents an efficient lightweight vehicle detection model, which is improved by using YOLOv4 network as a reference model. First, this paper uses CSPMobileViT network to replace the original backbone network, then replaces PANet with BiFPN, replaces 3×3 standard convolution in BiFPN with deep detachable convolution, and finally adds ECA module before BiFPN and YOLO-Head. In the loss function section, the Border Regression Loss CIoU is improved to Focal EIoU to solve the problem of difficult sample imbalance. The experimental results show that the mAP value of the improved network is 96.77%, the detection speed reaches 0.023 4 s per picture, the model size is only 32.76 MB, and the parameter amount is 8 587 541. Compared with the original algorithm, the mAP is improved by 1.54%, while the model size and number of parameters are only about 1/8 of the original model, and the FPS is improved by 7.5, so the improved algorithm has better detection effect.

Keywords: YOLOv4;MobileViT;vehicle detection;lightweight network

0 引 言

近些年来,随着人工智能的发展和图形计算硬件性能的提升,使得计算机视觉在无人驾驶领域被广泛的使用。在 2021 年 3 月 11 日,十三届全国人大四次会议表决通过了关于国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要的决议,在决议的数字化应用场景里提及了智能交通的重要性,而无人驾驶是智能交通的一个重要部分。因此在硬件性能受限的情况下进行快速准确的车辆检测是构建智能交通的必不可少部分[1]。

现在的车辆检测模型大概分为传统检测方法和深度学习目标检测方法两类。传统检测方法太过于依赖于人工设计的特征,所以人工设计的特征会对模型的性能起到至关重要的影响,而且通过滑动窗口来提取信息会存在时间复杂度和窗口冗余等问题,因此导致传统检测方法检测精度低、检测速度慢。而基于神经网络的深度学习目标检测方法是通过大量数据来学习得到特征,因此深度学习模型性能会优于传统检测方法模型,使得对深度学习进行研究的

收稿日期:2022-06-23
* 基金项目:国家自然科学基金(51468062)项目资助

学者越来越多，深度学习应用的范围越来越广。通常深度学习的目标检测模型大致可以分为两类，一类是以YOLO[2-5]为代表的一阶段检测方法，该类方法通常检测速度快，但是检测精度不高；另一类是以 R-CNN[6-8]为代表的二阶段检测方法，该类方法检测精度高，但是检测速度慢。因此对既检测速度快又有较好精度模型的研究是尤为必要的，尤其是在硬件性能受限的移动端，检测精度和检测速度都是轻量化模型所要面对的问题。

因此近些年来，国内外许多学者对此做出了相关的改进。厍向阳等[9]提出了基于残差网络的车辆检测方法实现了较高的检测精度，但检测速度较低难以达到实时性要求。宋焕生等[10]通过改进 Faster R-CNN 算法实现了对复杂场景下车辆目标检测，虽然基本能满足实时性要求，但模型对远场景下的小目标无法识别和检测。施培蓓等[11]为了解决行人检测的场景自适应的问题，对 RealAdaboost[12]框架进行了改进，提出了一种基于快速增量学习的行人检测方法。任丰仪等[13]将 YOLOv4 的主干网络替换为 MobileNet 改进后处理等方法，使得参数量只有原模型的一半，模型部署在 TX2 上 FPS 达到了 21.8，是原始的 YOLOv4 的 4.74 倍。涂媛雅等[14]针对轻量级车辆和行人检测网络做了进一步研究，虽然网络参数量大幅下降，但是模型的精度也不是太高。邓杰等[15]使用 YOLOv3 网络的基础上，提出时频域融合注意力模块 TFFAM，将频域通道注意力和空间注意力加入到网络中重新分配特征权重。陈玉敏等[16]提出基于时空融合加速的 Fast RCNN 运动车辆快速检测算法，不同于传统方法仅考虑运动车辆的单一特性，所提方法综合考虑运动车辆目标的时域运动特性和空域相关特性，但是该方法与一阶段检测算法相比要慢很多。

到目前为止，在硬件性能受限的嵌入式端进行车辆检测依然面临着挑战，对于模型的复杂性高、存储量大和检测速度慢的问题。本文所做的主要工作分为 4 个方面：

1）采用 YOLOv4 网络作为参考模型，使用带有 Transformer 结构的 MobileViT 网络来替换参考模型本身的主干网络 CSPDarknet53，并将 MobileViT 中的 MV2 block 残差结构改进为跨阶段部分模块 CSP 结构，在充分提取车辆特征信息的同时使得模型能够轻量化。

2）将 PANet 替换成 BiFPN，并且将 BiFPN 中的 3×3 标准卷积替换成深度可分离卷积，在大幅度降低模型的参数量的同时能更方便、快速的进行多尺度特征融合，同时提升模型对小目标的检查能力。

3）在 BiFPN 之前和 YOLO-Head 之前添加 ECA 模块，在不影响模型的检测速度的前提下，让模型在进行检测的同时能够关注车辆的位置信息，从而提高模型的检测精度。

4）改进损失函数，将用作边框回归损失 CIoU 的横纵比损失项拆分成预测的宽高分别与最小外接框宽高的差值，并且引入 Focal Loss，不仅加速了收敛提高了回归精度，还优化了样本不平衡问题。

1　轻量化 CSPMobileViT-YOLO 网络设计

1.1　YOLOv4 算法

1）基本结构

YOLOv4 模型于 2020 年 4 月被提出，该算法由主干网络 CSPDarknet53、SPP 结构[17]、PANet 特征融合结构[18]和 YOLO-Head 检测头四个部分组成，网络结构如图 1 所示。其预测处理过程如下：首先将 416×416 的图像通过 Mosaic 数据增强后传入目标检测网络，主干网络 CSPDarknet53 会对输入图片进行初步特征提取分别得到（52×52，26×26，13×13）3 种尺度的特征图；将 13×13 尺度的特征通过 SPP 的 3 次不同尺度的最大池化后进行拼接；然后将 3 种尺度特征通过 PANet 进行特征融合，将融合后的特征进行特征提取来获得更深层次具有语义信息的特征，将 3 种尺度的深层次特征图输入到 YOLO-Head；最后通过 YOLO-Head 分别对这 3 种不同尺度的特征图进行分类回归预测结果。

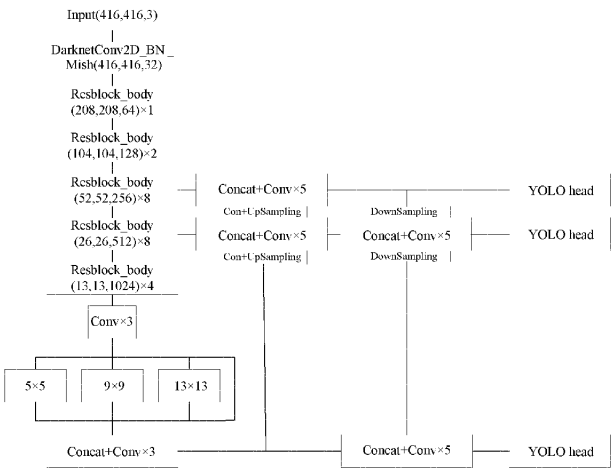


图 1　YOLOv4 网络结构

2）损失函数

YOLOv4 的损失函数由边框回归损失、置信度损失和目标分类损失三个部分构成。和上一代 YOLOv3 的损失函数相比，YOLOv4 的损失函数把边框回归损失部分从 MSE 改进为 CIoU 损失。CIoU 损失是在 IoU 的基础上不仅考虑了边框的重合度和中心距离，还考虑了宽高比的尺度信息，如式(1)所示。

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

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

$$v = \frac{4}{\pi^2}\left(\arctan\left(\frac{w^{gt}}{h^{gt}}\right) - \arctan\left(\frac{w}{h}\right)\right)^2 \tag{3}$$

$$\alpha = \frac{v}{1 - IoU + v} \tag{4}$$

式中:IoU 代表预测框和真实框的交并比,b 表示预测框,b^{gt} 表示真实框,$\rho(b,b^{gt})$ 表示预测框和真实框各自中心点之间的欧式距离,c 表示刚好包含预测框和真实框的最小框的对角线距离,v 表示测量长宽比的一致性,α 是一个平衡参数,w^{gt} 和 h^{gt} 为真实框的宽、高,w 和 h 为预测框的宽、高。

3)K-means 聚类

在目标检测中,为了使得模型能有更好的检测效果,通常会提前标定锚点框对目标进行聚类。YOLOv4 算法使用 K-means 聚类算法对训练集中的目标框进行聚类,该算法把每个对象与聚类中心的欧氏距离当作衡量相似度的指标,通过将每个对象点归到最相似的类中,再来计算每个类的聚类中心,重复此过程直至结果不再改变。

针对本文采用的数据集,将 K-means 算法的 K 值设为 9,再通过聚类算法计算出锚框的尺度分别为(38, 29)、(76, 48)、(127, 69)、(194, 95)、(121, 170)、(228, 163)、(302, 159)、(393, 189)、(347, 339)。图 2 为本实验 K-means 聚类效果图。

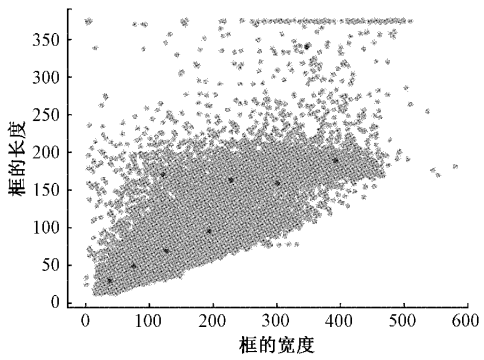


图 2　K-means 聚类效果图

1.2　CSPMobileViT-YOLO 优化模型

1)主干网络优化

虽然 CSPDarknet53 对特征初步提取很有效果,但是该网络属于大型网络,在硬件受限的条件下存在部署困难的问题,因此本文选用 MobileViT 模型[19]来替换 YOLOv4 的主干网络去对特征进行初步提取。MobileViT 模型是一种用于移动设备的轻量级通用视觉 Transformer。

MobileViT 模型的核心思想就是通过 Transformer as Convolution 学习全局表达能力,这样可以使得该模型能够同时具备 Convolution 和 Transformer 两者各自的优点。在参数量相当的情况下,与当前的轻量级 CNN 模型相比,MobileViT 模型在不同的移动端视觉任务中均展现了更好的性能。与使用数据增强的 CNN 模型相比,MobileViT 模型也展现出更好的泛化能力。图 3 为 MobileViT 模型结构图,其预测处理过程如下:将输入图像后接 3×3 标准卷积,并做 2 倍下采样;之后通过 5 个 MV2 block,其中 stride=1 的 MV block 进行特征提取,stride=2 的 MV block 做 2 倍下采样;将得到的特征图间隔传入 MobileVit block 和 MV2 block stride=2;然后使用 1×1 标准卷积进行通道压缩;最后进行全局平均池化来获取预测结果。

MobileViT block 的具体结构如图 3 所示,将输入 $H\times W\times C$ 的特征图通过 $n\times n$ 标准卷积和 1×1 标准卷积放缩通道数为 d,得到 $H\times W\times d$ 的特征图。其中 $n\times n$ 标准卷积编码原特征图的局部空间信息,1×1 标准卷积用于升维。然后将获得的 $H\times W\times d$ 的特征图展开为 $P\times N\times d$,然后将其输入至 Transformer 去提取特征图的全局空间信息,把输出的 $P\times N\times d$ 特征图再折叠复原至 $H\times W\times d$,这里的 $P=H\times W$,$N=H\times W/P$。最后将 $H\times W\times d$ 特征图通过 1×1 标准卷积复原回 $H\times W\times C$,再与最开始的

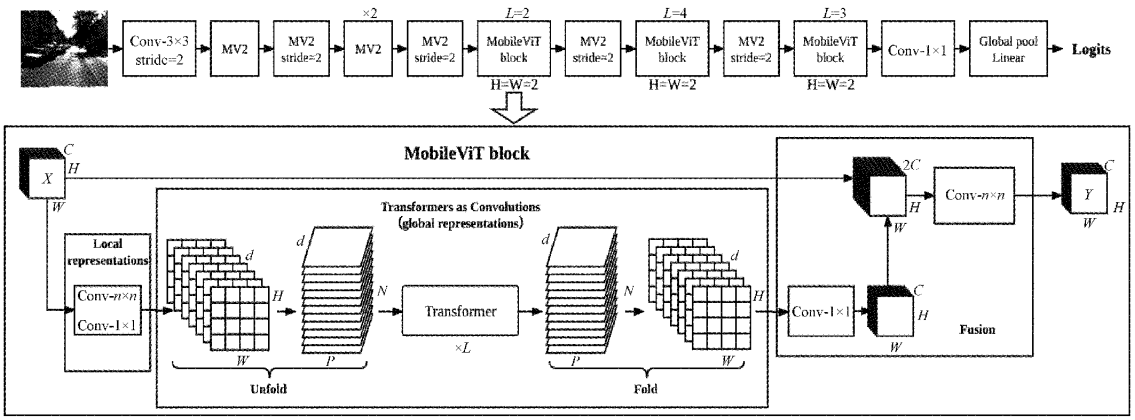


图 3　MobileViT 网络结构

输入 $H\times W\times C$ 的特征图做类似于 shortcut 的拼接,使用 $n\times n$ 标准卷积做通道融合得到最终的 $H\times W\times C$ 输出特征图。

在一个普通的 ViT 模型中,将输入 $H\times W\times C$ 的特征图压扁为 $N\times PC$ 的特征图,再将其进行序列变换到 d 维空间得到 $N\times d$ 的特征图,对该特征图进行位置编码,然后输入 Transformer,Transformer 输出的特征图经过线性变化,得到最终结果。与普通 ViT 模型相比,普通 ViT 模型的 $N\times d$ 的特征图丧失了每个像素点的空间顺序,但是 MobileViT 增加了 1 维得到 $P\times N\times d$ 特征图,使得能够

保留每个 patch 中每个像素点的位置。使用普通卷积的神经网络有局部连接这个重要特点，并且通过感受野来描述神经元所看到的输入区域有多大，而对于 MobileVit 模型，在 Transformer 之后输出的特征图，每个像素点都来源于输出特征图的所有像素点。如图 4 所示，原始特征图被分成了 9 个 patch，最中间黑色像素点与 9 个 patch 中黑色像素点做 self-attention，而每个 patch 中的黑色像素点都通过卷积获得它周围灰色像素点的局部分析。这相当于黑色像素包含了输入特征图所有像素的信息，具有全局的感受野。并且最后通过 Transformer 输出的特征图的每个像素点都包含输入特征图所有像素的信息。

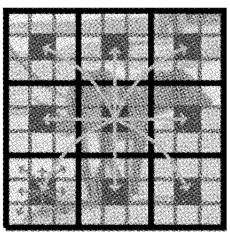


图 4　感受野示意图

MV2 block 分为两种形式，stride＝1 和 stride＝2，如图 5(a)、(b)所示，stride 控制特征图的大小。block 开始的 1×1 标准卷积用来调整通道数，后面的 1×1 标准卷积一般是将通道数量调回来。现有的模型经过轻量化后，会大大降低模型对特征的提取能力，从而降低准确性，为了增强轻量化模块对特征提取的能力，本文对 MV2 stride＝1 进行改进，通过增加跨阶段部分模块 CSP 结构，使得在增强模块学习能力的同时，既保持了轻量化的准确性，又下降了计算瓶颈，还下降了内存成本，改进后的 CSPMV2 模块如图 5(c)所示。

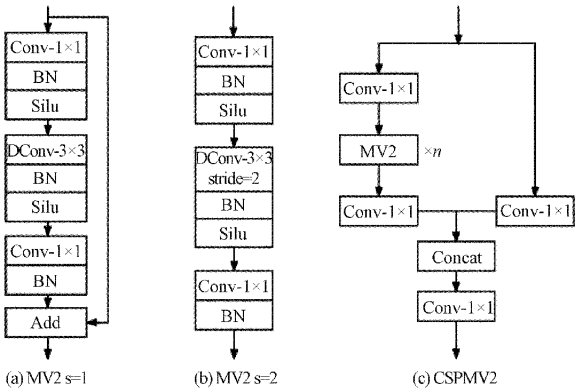


图 5　MV2 block 和改进后的 CSPMV2

在通过主干网络获取的特征图时，网络中越靠前的网络结构越是提取形状等低级特征，越是靠后的网络结构越是提取语义等高级特征。而针对本文检测对象，并不需要涉及太复杂的语义信息，并且过深的网络结构可能会带来网络退化和梯度不稳定等问题，这些问题会使得性能反而开始下降，所以本文将 MobileViT 模型的第十层 32 倍下采样后的网络全部舍弃，再将剩余的部分作为主干网络接入 SPP 结构和 BiFPN 结构，改进后的 CSPMobileViT-YOLO 网络结构如图 6 所示。

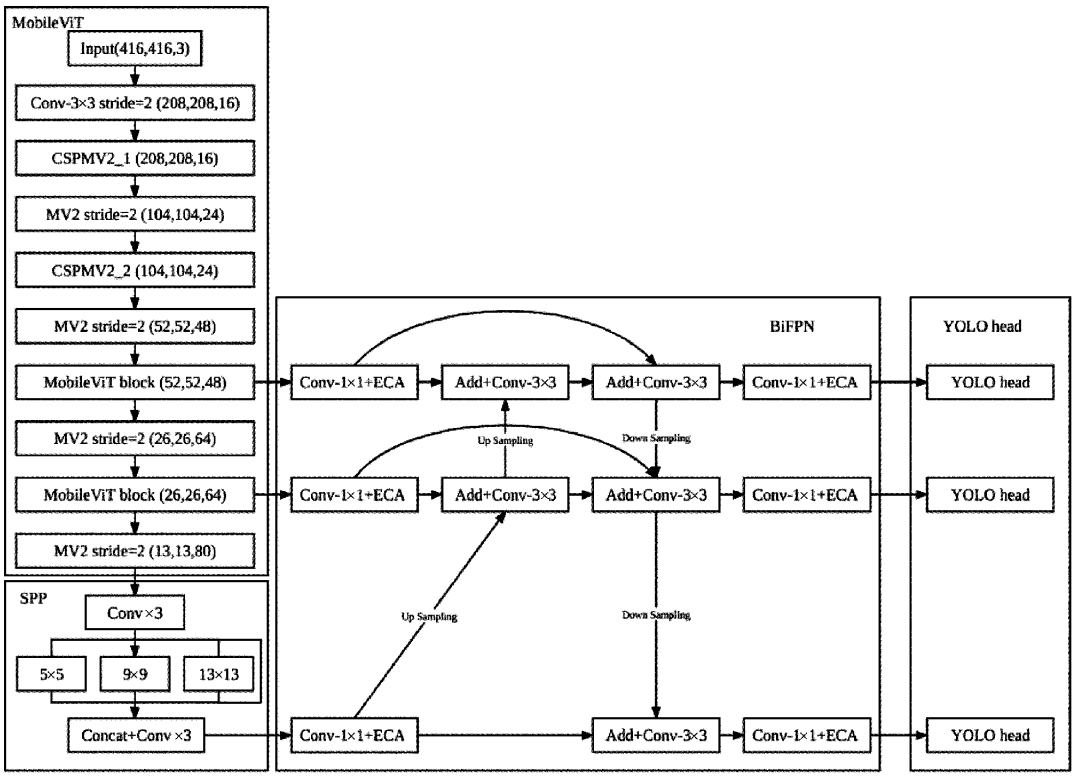
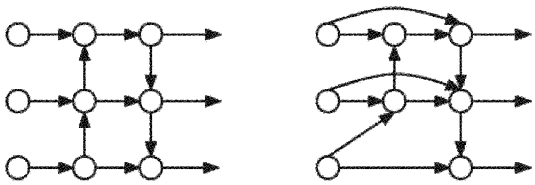


图 6　CSPMobileViT-YOLO 网络结构

2)特征加强网络优化

YOLOv4 原始的特征加强网络是 PANet,网络结构示意图如图 7(a)所示,它是在 FPN 结构的从上到下的特征融合路径的基础上,新增了一条从下到上的融合路径,进而实现不同特征层之间的特征融合。本文采用一种加权双向特征金字塔模型 BiFPN 来替换 PANet,网络结构示意图如图 7(b)所示。与 PANet 结构相比,BiFPN 结构删除了只有一个特征输入的节点,因为一个节点如果只有一个特征输入,那么它既无法进行特征融合又增加了模型的参数量,所以这种节点对融合不同特征的特征网络来说贡献就会更小;BiFPN 结构添加了不相邻输入节点和输出节点之间的跳跃连接,因为这些节点在同一层,所以这些节点的特征图大小都相同,在不增加太多计算成本的同时,可以融合更多的特征。



(a) PANet结构示意图 (b) BiFPN结构示意图

图 7 PANet 和 BiFPN 结构示意图

BiFPN 结构的参数量主要集中 3×3 标准卷积和上、下采样块中,为了能降低模型的参数量和计算量来提升 CSPMobileViT-YOLO 的运行效率,本实验采用更有效率的深度可分离卷积来替换 BiFPN 结构中所有的 3×3 标准卷积。

深度可分离卷积是由 Depthwise 和 Pointwise 两个部分结合起来。与标准卷积相比,在对同一输入获得相同大小的输出的情况下,深度可分离卷积的参数数量和运算成本比较低。例如在输入和 12×12×18 和输出为 8×8×256 的情况下,标准卷积的参数量为:5×5×128×256=819 200,标准卷积的计算量为:5×5×128×8×8×256=52 428 800;深度可分离卷积的参数量为:5×5×1×128+1×1×128×256=35 968,深度可分离卷积的计算量为:5×5×1×8×8×128+1×1×128×8×8×256=2 301 952。从参数量和计算量的对比可以看出深度可分离卷积的效率更高。

3)添加 ECA 注意力模块

ECA 注意力模块[20]是一种具有高效相关性的非常轻量级的模块,它主要由通过非线性自适应确定的一维卷积组成。其结构如图 8 所示。在 SE 注意力模块中,SE 注意力模块会对特征进行通道压缩,而这种方法会损失通道之间的依赖关系,从而给模型带来不利的影响。于是在此观点上,ECA 注意力模块通过采用 1 维标准卷积来避免降维,在高效的同时实现了局部跨通道交互。具体步骤如下:

(1)将 $H\times W\times C$ 的输入特征图进行全局平均池化操作压缩成 $1\times 1\times C$;

(2)由式(5)确定 k 的大小,进行卷积核大小为 k 的 1 维标准卷积操作,并经过 Sigmoid 激活函数得到各个通道的权重 w;

(3)将 $1\times 1\times C$ 的权重与 $H\times W\times C$ 的原始输入特征图对应元素相乘,得到最终 $H\times W\times C$ 的输出特征图。

$$k=\left|\frac{\log_2(C)}{r}+\frac{b}{r}\right| \tag{5}$$

式中:$r=2$、$b=1$,C 为通道数,k 为计算出来临近的奇数值。

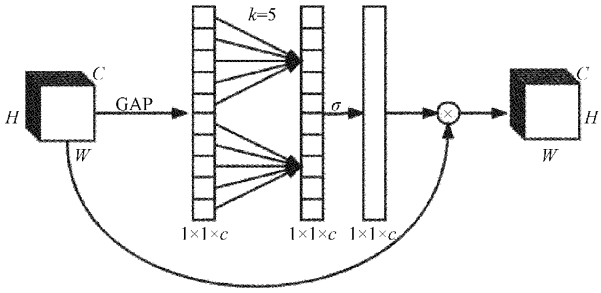


图 8 ECA 注意力模块

可以看出,ECA 注意力机制只需要确定一个参数 k 来决定 1 维卷积的大小,相比其他注意力机制来说,由于 ECA 注意力机制对原网络的改变最小,运算也简便,所以对原网络运行速度的影响也是最小。

4)Focal EIoU

在损失函数方面,YOLOv4 使用考虑了边框的重合度、中心距离和长宽比的尺度信息的 CIoU,但是通过 CIoU 的公式中的 v 反映的长宽比的差异,而不是长宽分别与其置信度的真实差异,当采用 CIoU 作为损失时可能给模型有效优化带来负作用。对于这个问题,在 CIoU 的基础上把长宽比拆开,并且通过引入 Focal 聚焦更优质的锚框,如式(7)所示。

$$EIoU=1-IoU+\frac{\rho^2(b,b^{gt})}{c^2}+\frac{\rho^2(w,w^{gt})}{Cw^2}+\frac{\rho^2(h,h^{gt})}{Ch^2} \tag{6}$$

$$Focal\_EIoU=IoU^{\gamma}\times EIoU \tag{7}$$

EIoU 损失函数是由边框的重合度损失、中心距离损失和长宽损失 3 个部分构成。边框的重合度损失和中心距离损失延续 CIoU 中的方法,为了使得收敛速度更快,长宽损失直接使目标框与锚框的宽度和高度之差最小。其中 Cw 是覆盖两个框的最小外接框的宽度,Ch 是覆盖两个框的最小外接框的高度。为了解决训练样本不平衡的问题,在 EIoU 的基础上结合 Focal Loss 提出一种 Focal EIoU Loss。γ 表示控制异常值抑制程度的参数,当预测框和真实框的 IoU 越大时,就给损失一个更大的权值,从而使得给回归一个更大的损失,这样能够提高模型训练时的

回归速度和精度。

2　实验分析与结果

2.1　实验配置

为验证算法的优越性，建立针对数据集的对比实验和消融实验，实验软硬件环境配置如表 1 所示。

表 1　软硬件环境配置表

	配置版本
操作系统	64 位 Windows10
处理器	Intel(R) Xeon(R) w_2223
GPU	RTX 2080Ti×2
内存	32 GB
Cuda	Cuda 10.1
Python	Python 3.7
深度学习框架	Pytorch 1.7.0
开发工具	PyCharm 2020.1.2

2.2　数据集

本文采用 KITTI 数据集[21]作为实验数据集，该数据集是由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合制作，是目前自动驾驶场景中最大的计算机视觉算法评估数据集。该数据集图像具有不同程度的截断和遮挡[22]，并且单张图像最多拥有 15 辆待测车辆。根据本文的车辆检测任务，从数据集的类别中选取 truck、van 和 car 共 3 类来作为待检测类别。该数据集包含 7 481 张图像，根据 8∶1∶1 的比例将数据集分成训练集、验证集和测试集。数据集随机划分后的详细情况如下表 2 所示。

表 2　数据集划分详情表

类别	数量	Car	Van	Truck
训练集	5 984	22 701	2 991	852
验证集	748	2 924	312	114
测试集	749	3 127	311	128

2.3　评价指标

为了验证改进后的网络模型 CSPMobileViT-YOLO 的性能，本文采用模型大小(Size)、参数量(Param)、平均精准度(mAP)和每秒检测帧数(FPS)四个指标对算法进行评估。平均精度和每秒检测帧数如式(10)、(11)所示。

$$P=\frac{TP}{TP+FP} \tag{8}$$

$$AP=\frac{\sum_{i-1}^{Sum}P_i}{Sum} \tag{9}$$

$$mAP=\frac{\sum_{i=1}^{Class}AP_i}{Class} \tag{10}$$

$$FPS=\frac{fn}{T} \tag{11}$$

其中，TP 表示检测结果为正的正样本，FP 表示检测结果为正的负样本，$TP+FP$ 表示检测结果为正的样本总数，P 表示检测结果为正的样本中真正正样本占的比例，Sum 表示测试集中的图片总数，$Class$ 表示检测类别数。fn 表示模型处理图像的总数，T 表示所用时间。

2.4　模型训练

本文使用通过冻结和解冻两个阶段进行训练模型。将标签平滑 *label_smoothing* 设置为 0.01，在冻结阶段设置训练轮数 *Freeze_Epoch* 为 20，冻结阶段学习率 *Freeze_lr* 为 0.001；在解冻阶段设置训练轮数 *Unfreeze_Epoch* 为 80，解冻阶段学习率 *Unfreeze_lr* 为 0.000 1。设置权值衰减 *weight_decay* 为 0.000 5，动量 *momentum* 为 0.937。

2.5　实验结果分析

为了验证 Focal EIoU 损失函数的有效性，本文将初始 YOLOv4 网络结构中的 CIoU 损失函数分别换为 EIoU 和 Focal EIoU 损失函数，实验结果如表 3 所示。

表 3　损失函数性能对比

损失函数	mAP/%
YOLOv4＋CIoU	94.1
YOLOv4＋EIoU	94.8
YOLOv4＋Focal EIoU	95.2

实验结果表明，采用 Focal EIoU 作为定位回归损失函数相比 CIoU，mAP 提高了 1.1%，因此将 CIoU 上的宽高比拆开，更加有利于模型的训练，用 Focal EIoU 作为边界框损失衡量标准可以提高检测的精度。

为了能更直观的展现本文改进的算法的性能，图 9 列举了本文设计的轻量化模型在图片运行结果。可以从结果图中看出，本文设计的轻量化的模型能够有效地对图像、视频中所包含不同大小车辆的位置进行检测。与原 YOLOv4 算法相比，检测结果有些许提升，并且实时性更佳，模型更轻量化更适合在嵌入式端进行部署。

本文将改进后的模型与近些年的目标检测模型进行对比，采用 mAP、Size、Param 和 FPS 为性能指标，在相同的环境相同的数据集上进行性能对比试验。对比结果如表 4 所示。

从表 4 中可以看出，本文改进的 CSPMobileViT-YOLO 算法在保证实时检测的情况下，检测精度也高于其他算法。图 10 为 AP 变化和 mAP 变化图，truck、van 和 car 的 AP 值分别为 97.82%、96.96%、95.54%。其中本文改进算法的 mAP 为 96.77%，与 YOLOv3 相比 mAP 提高了 8.51%，与原模型 YOLOv4 相比 mAP 提高了 1.54%，与其他轻量化网络相比 mAP 也有不小的提高，并且 FPS 可达到 42.66。与文献[23]所提出的模型相比，在

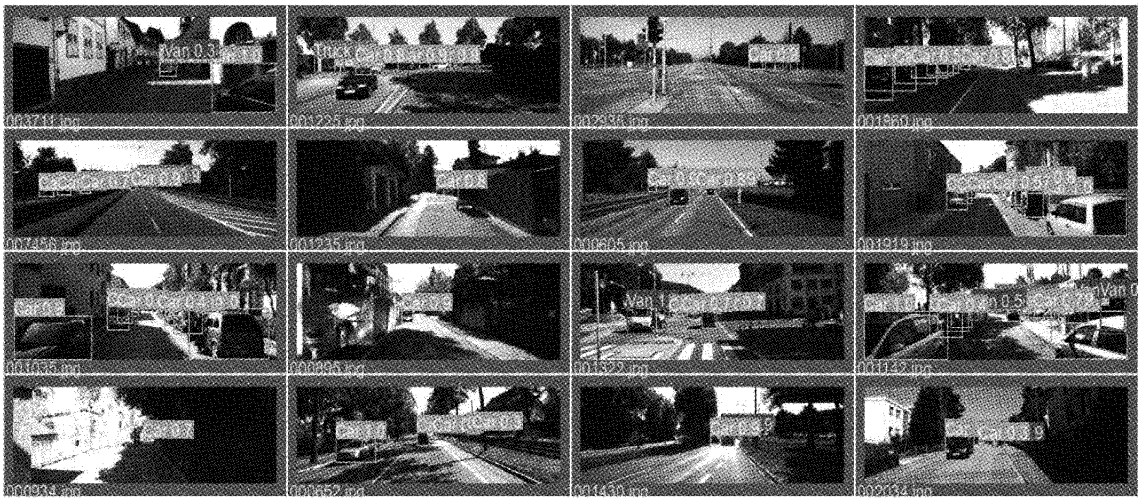
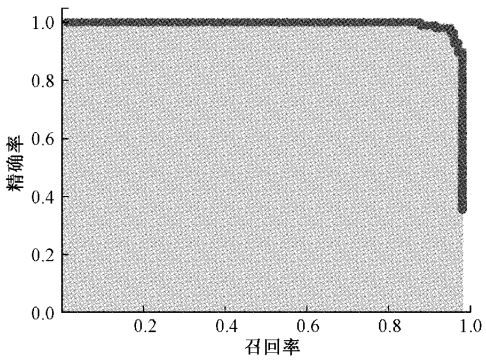


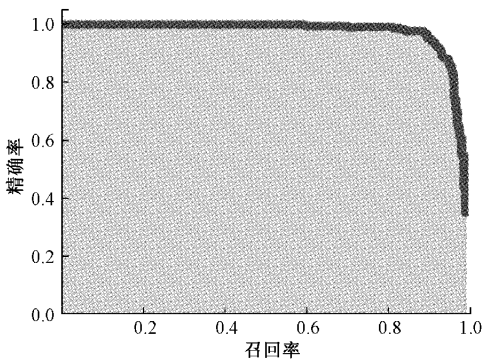
图 9　模型运行结果图

表 4　不同目标检测算法的对比实验

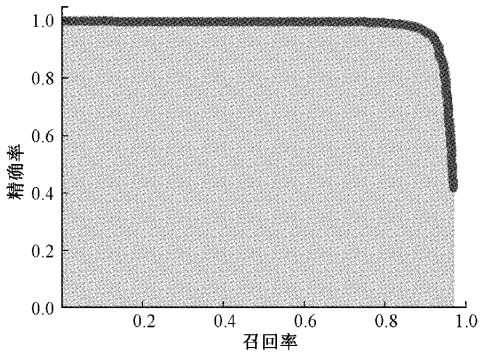
模型	mAP/%	Size/MB	Param/M	FPS
YOLOv3	88.26	236.32	61.9	24.30
YOLOv4	95.23	245.53	64.4	35.16
YOLOv4-Tiny	62.32	23.62	6.1	81.12
Mobilenetv1-YOLOv4	86.68	156.22	40.9	38.24
Mobilenetv2-YOLOv4	87.12	149.01	39.1	37.78
Mobilenetv3-YOLOv4	89.87	152.55	40.0	40.34
Ours	96.77	31.18	8.2	42.66



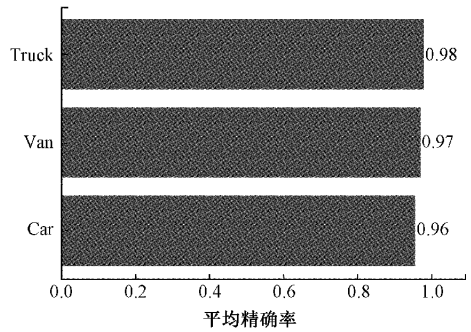
(a) Truck AP变化图



(b) Van AP变化图



(c) Car AP变化图



(d) mAP总图

图 10　AP 和 mAP 变化图

略微降低精确度的情况下,大大降低了模型的大小和参数量,并且提升了模型检测速度。总的来看,本文提出的算法对主干网络进行改进,在减小参数和复杂度的同时,同时也拥有不错的特征提取能力。同时在特征融合网络中引入注意力模块,使得即便在轻量化的过程中也不会损失检测精度,提高对特征图的融合能力。本文改进算法在保证实时性的情况下,有效地提高了对车辆特征的提取能力,为嵌入式移动端部署提供了可行性。

2.6　消融实验

本文改进算法是以 YOLOv4 作为基础模型,在网络结

构提出了以下改进：使用 CSPMobileViT 网络来作为主干网络；使用 BiFPN 作为特征融合网络；使用深度可分离卷积来替换 3×3 标准卷积；加入 ECA 注意力模块。为突出本文算法中四处改进点的有效性，分别对这几处改进点做消融实验，其中消融实验的实验配置、参数设置与本文相同，消融实验结果如表 5 所示。

表 5　消融实验对比

模块 / 模型	CSPMobileViT	DW	ECA	BiFPN	mAP/%	Size/MB	Param/M
原模型					95.32	245.53	64.4
	√				95.87	146.54	38.4
	√	√			86.74	38.74	10.2
	√	√	√		91.28	38.74	10.2
	√	√	√	√	96.77	31.18	8.2

由表 5 可看出，首先从平均精度 AP 上进行分析：由于 ECA 注意力模块可提升车辆重要特征信息的提取能力和获得该目标物体的位置信息，BiFPN 结构可以在不增加大量参数的情况下融合更多特征，Focal EIoU Loss 将宽高损失拆分，使得模型收敛更快，同时也考虑样本不平衡的问题，从而使本文改进算法的 AP 值在此实验算法中最高。并且本文算法使用了许多轻量化方法，其中采用 MobileViT 来作为主干网络最为关键，因为其兼具 Convolution 和 Transformer 的优点，在降低模型大小和参数量的同时，通过 Transformer 的每个像素点都来源于输出特征图的所有像素点，即每个像素点都拥有全局信息，使得模型能保持较好的检测性能。

3　结　　论

本文提出的 CSPMobileViT-YOLO 实时车辆检测轻量化模型，首先将 YOLOv4 的主干网络 CSPDarknet53 改进为的 CSPMobileViT，并且将特征融合网络替换为 BiFPN。接下来将优化 CSPMobileViT-YOLO 模型，通过添加卷积注意力模块 ECA 提高了卷积神经网络输出特征图的全局特征；其次通过引入 Focal EIoU Loss 有效地降低了样本不平衡的问题。实验结果表明，本文模型由于主干网络中的 Transformer 模块，与其他轻量化模型相比有着更好的检测精度，与使用数据增强的模型相比，本文模型有更好的泛化性，并且本文模型能更好的满足车辆检测任务的实时性要求。这有利于将算法部署到计算能力和内存等资源有限的无人嵌入式平台上，使得无人驾驶汽车能够对视野中的目标进行实时识别，提高了无人驾驶汽车的场景理解能力。今后的研究工作中，在此模型的研究基础上，将会尝试模型剪枝、参数量化、知识蒸馏等手段，继续尝试轻量化处理，从而提升模型的检测速度。

参考文献

[1] 陈立潮，王彦苏，曹建芳. 基于 Dense-YOLOv3 的车型检测模型[J]. 计算机系统应用，2020，29(10)：158-166.

[2] REDMON J，DIVVALA S，GIRSHICK R，et al. You only look once：Unified，real-time object detection[C]. IEEE Conference on Computer Vision & Pattern Recognition，IEEE，2016：779-788，DOI：10.1109/CVPR.2016.91.

[3] REDMON J，FARHADI A. YOLO9000：Better，faster，stronger[C]. IEEE Conference on Computer Vision & Pattern Recognition，IEEE，2017：6517-6525，DOI：10.1109/CVPR.2017.690.

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

[5] BOCHKOVSKIY A，WANG C Y，LIAO H. YOLOv4：Optimal speed and accuracy of object detection[J]. ArXiv Preprint，2020，ArXiv：2004.10934.

[6] GIRSHICK R，DONAHUE J，DARRELT，et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. IEEE Conference on Computer Vision & Pattern Recognition，IEEE，2014：580-587，DOI：10.1109/CVPR.2014.81.

[7] GIRSHICK R. Fast R-CNN[C]. IEEE Conference on Computer Vision & Pattern Recognition，IEEE，2015：1440-1448，DOI：10.1109/ICCV.2015.169.

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

[9] 厍向阳，韩伊娜. 基于残差网络的小型车辆目标检测算法[J]. 计算机应用研究，2020，37(8)：2556-2560.

[10] 宋焕生，张向清，郑宝峰，等. 基于深度学习方法的复杂场景下车辆目标检测[J]. 计算机应用研究，2018，35(4)：1270-1273.

[11] 施培蓓，刘贵全，汪中. 基于快速增量学习的行人检测方法[J]. 小型微型计算机系统，2015，36(8)：1837-1841.

[12] SCHAPIRE R E，SINGER Y. Improved boosting algorithms using confidence-rated predictions[J]. Machine Learning，1999，37(3)：297-336.

[13] 任丰仪，裴信彪，乔正，等. 融合 CBAM 的 YOLOv4 轻量化检测方法[J/OL]. 小型微型计算机系统：1-8[2022-06-03]. http://kns.cnki.net/kcms/detail/21.1106.tp.20220301.0935.002.html.

[14] 涂媛雅,汤国放,张建勋. Lite-YOLOv3 轻量级行人与车辆检测网络[J/OL]. 小型微型计算机系统: 1-8 [2022-06-03]. http://kns.cnki.net/kcms/detail/21.1106.tp.20211015.0001.002.html.

[15] 邓杰,万旺根. 基于改进 YOLOv3 的密集行人检测[J]. 电子测量技术,2021,44(11):90-95.

[16] 陈玉敏,李淼,房晓丽. 基于时空融合加速的 Fast RCNN 运动车辆检测算法[J]. 电子测量技术,2020,43(3):139-145.

[17] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[18] WANG K, LIEW J H, ZOU Y, et al. Panet: Few-shot image semantic segmentation with prototype alignment[C]. IEEE/CVF International Conference on Computer Vision, 2019: 9197-9206, DOI: 10.1109/ICCV.2019.00929.

[19] MEHTA S, RASTEGARI M. MobileVit: light weight general purpose, and mobile friendly vision transformer [J]. ArXiv Preprint, 2021, ArXiv: 2110.02178.

[20] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), IEEE, 2020: 11531-11539, DOI: 10.1109/CVPR42600.2020.01155.

[21] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset [J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[22] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012: 3354-3361, DOI: 10.1109/CVPR.2012.6248074.

[23] 熊李艳,涂所成,黄晓辉,等. 基于 MobileVit 轻量化网络的车辆检测方法[J]. 计算机应用研究, 2022, 39(8):5.

作者简介

黄德启,副教授,硕士生导师,主要研究方向为智能交通。
E-mail:dqhuang88@qq.com

郑玉珩,硕士研究生,主要研究方向为深度学习、图像处理、智能交通。
E-mail:1812294492@qq.com