第 58 卷 第 16 期/2021 年 8 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于改进 YOLOv3 的车辆识别方法

王永顺，贾文杰*，王晨飞，宋慧

兰州交通大学电子与信息工程学院，甘肃 兰州 730070

摘要 针对车辆目标检测中存在的小目标检测准确率低、系统鲁棒性差的问题，提出一种改进的 YOLOv3 算法对车辆进行目标检测。首先，该算法将空洞卷积引入到 YOLOv3 算法的下采样层，提高特征图的分辨率，加强对小目标的检测效果；其次，针对车辆图像中小目标识别的问题，将 YOLOv3 的 3 个检测尺度增加至 4 个并相互融合不同尺度特征层的信息，改进后的空间金字塔结构实现了对小目标检测进行进一步增强的目标；最后，采用 Complete IoU (CIoU)作为损失函数，使目标框回归更加稳定，并且训练中不会出现发散现象。在 KITTI 数据集上的测试结果表明，所提改进的 YOLOv3 算法能获得较高的检测精度，平均检测精确度提高了 4.6%，检测速度约为 44.1 frame/s，在提高精度的前提下仍保持良好的检测速率。

关键词 图像处理；车辆目标检测；YOLOv3；空洞卷积；尺度检测；损失函数

中图分类号 TP183 **文献标志码** A **doi**：10.3788/LOP202158.1610010

Vehicle Recognition Method Based on Improved YOLOv3 Algorithm

Wang Yongshun, Jia Wenjie*, Wang Chenfei, Song Hui

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu 730070, *China*

Abstract An improved YOLOv3 algorithm that detects target vehicles is proposed to address the problems of low detection accuracy of small targets and poor robustness of systems in target vehicle detection. First, the proposed algorithm introduces the dilated convolution into the downsampling layer of the YOLOv3 algorithm, improving the resolution of the feature maps and detection effect of small targets. Second, to address the problem of small target recognition in vehicle images, the proposed algorithm increases the three detection scales of YOLOv3 to four in addition to connecting and fusing the information with different scales, and the improved feature pyramid structure further improves small target detection. Finally, using Complete IoU (CIoU) as the loss function makes the target frame regression more stable, and there is no divergence in training. The KITTI dataset test results show that the improved YOLOv3 algorithm can achieve high detection accuracy. The proposed algorithm improves the average detection accuracy by 4.6%, and the detection rate is approximately 44.1 frame/s. On the premise of improving the accuracy, the proposed algorithm maintains a high detection rate.

Key words image processing; vehicle target detection; YOLOv3; dilated convolution; scale detection; loss function

OCIS codes 100.4997; 100.3008; 110.2970

1 引 言

近年来，随着我国车辆数目的飞速增长，车辆违规行为不断增加。对交通图像或视频中的车辆进行快速而又准确的检测成为城市交通管理的重要工作。在大量的检测算法中，基于深度卷积神经网络的算法引起了极大的关注并被广泛使用。因此对复杂场景中小尺度和不同类型的车辆进行准确并实时的检测是最具挑战性的任务[1-2]。

目前，一些具有针对性的车辆数据集已经被提出，如 KITTI 数据集[3]、BDD100K 数据集[4]、CamVid 数据集[5]等，这些数据集可用于评价各种

收稿日期：2020-11-17；**修回日期**：2020-12-04；**录用日期**：2020-12-17

基金项目：国家自然科学基金(61366006)

通信作者：*361658066@qq.com

检测算法的性能。在深度学习中，大量关于卷积神经网络的方法被用于分类任务[6-7]，这些方法在目标检测任务中[8]也取得了广泛成功。这些方法可分为两大类，基于区域的方法和基于回归的方法。Girshick 等[9]提出区域卷积神经网络(RCNN)目标检测算法，RCNN 采用卷积神经网络来提取图像特征，使用选择性搜索方法对建议框进行优化。这在传统机器学习算法提取目标框的基础上有了很大的提升，但 RCNN 有输入尺寸固定、训练过程繁杂、测试速度较慢的缺陷，因此检测效率较低，检测速率仅为 0.02 frame/s，平均检测准确率(mAP)为 66.0%。He 等[10]在 RCNN 的基础上提出了空间金字塔池化网络(SPP-Net)。SPP-Net 对固定输入层尺寸的问题给出了解决方法，但在反向传播阶段，只更新金字塔池化层后的全连接层，无法更好地反向传播误差，其训练数据还需存储在磁盘中。SPP-Net 的 mAP 达到 63.1%，检测速率为 0.43 frame/s，效率依旧低下。Girshick 等对 RCNN 进行改进，提出了 Fast RCNN[11]和 Faster RCNN[12]。Fast RCNN 在 RCNN 的基础上又吸收了 SPP-Net 的特点，很好地优化了模型，使得测试速度更快，训练空间更小，其 mAP 为 66.9%，检测速率为 0.5 frame/s。但 Fast RCNN 仍然使用选择性搜索算法来提取候选框，这无法达到实时检测的要求，且不符合端到端的训练与调试。Faster RCNN 是在 Fast RCNN 基础上的区域生成网络(RPN)，检测图像中的区域是否存在目标。与 Fast RCNN 相比，Faster RCNN 真正实现了端到端的检测，在检测速度上有了很大的提升，其 mAP 为 73.2%，检测速率为 5 frame/s，但是离实时检测仍有差距。Dai 等[13]提出了区域全卷积网络(R-FCN)，R-FCN 采用深度残差网络进行改进，在检测速度上有了进一步的提升，mAP 为 77.6%，检测速率为 6 frame/s，但还是无法达到实时检测的目标。

为了实现检测速度和精度的平衡，提出了基于回归的检测方法。这类方法可以直接在对目标物体进行位置回归计算的同时给出物体分类得分。Redmon 等[14]提出了 you only look once(YOLO)检测算法，YOLO 使用回归算法对物体进行检测，mAP 为 63.4%，检测速度达到 45 frame/s，但是存在检测精度不高和位置定位误差大等缺点。Liu 等[15]提出单步多框检测器(SSD)，SSD 通过融合多个卷积层的特征图来增强模型特征的表达能力，mAP 为 76.8%，检测速率为 19 frame/s。SSD 在确保检测速度快的同时，检测精度相较 YOLO 也有了提升。但 SSD 中候选框的选择和特征图大小的调试都需要人工设置，太过依赖经验。Redmon 等在 YOLO 模型的基础上提出了 YOLOv2[16]和 YOLOv3[17]，YOLOv2 的检测速度达到 67 frame/s，mAP 为 76.8%。YOLOv3 在 YOLOv2 的基础上结合了残差网络和特征金字塔结构(FPN)[18]，加强了回归检测方法对小目标的检测精度。在 COCO 数据集上进行测试，YOLOv3 的 mAP 达到 57.9%，识别准确率相比基于区域的检测方法平均高出 3%，检测速率相比 SSD 也提高了 3 倍。但 YOLOv3 依然存在对小目标的漏检和误检情况，检测准确率仍有待提升。

本文对 YOLOv3 进行改进，首先将 YOLOv3 中 3 个检测尺度扩展为 4 个，以增强对浅层特征图中小目标的识别；其次将空洞卷积运用于下采样层来扩大感受野，从而提取到更多的图像细节信息；最后采用 Complete IoU(CIoU)损失函数[19]，使预选框的回归更加稳定，精度更高。

2 改进的 YOLOv3 算法

2.1 YOLOv3 原理和相关改进工作

YOLOv3 是 YOLO 的第三个版本，由 Joseph Redmon 和 Ali Farhadi 二人提出。该模型将 Darknet-53 网络作为特征提取器，主要借鉴了残差网络(ResNet)的做法，在层与层之间引入直连结构，适当地跳过卷积。这解决了深度卷积神经网络出现梯度消失的问题，并能更好地提取图像特征。YOLOv1 和 YOLOv2 均使用单尺度检测方法。单尺度检测网络如图 1 所示，图像经过多层卷积后，只对最后特征层作出预测。此方法经过下采样后忽略了图像细节信息，仅具有弱特征，不利于边框的回归和小目标的识别。

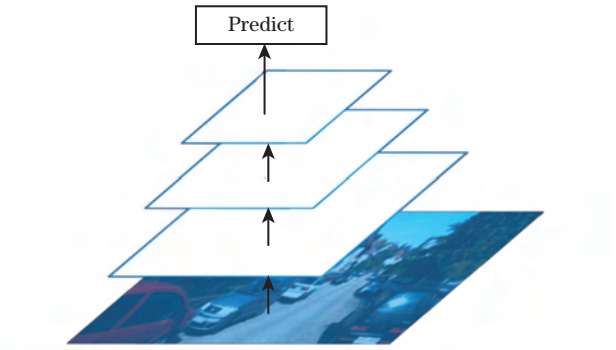

图 1 单尺度检测网络

Fig. 1 Single-scale detection network

YOLOv3 使用 FPN 的多尺度检测方法，将图片转换成三个不同尺度的特征图来检测大、中、小三类物体。FPN 的结构如图 2 所示，3 个检测的尺度大小分别为 13×13、26×26、52×52。对 3 个尺度的特征层进行特征融合处理，利用多尺度的特征增加特征的丰富程度，既保留了深层特征图的语义信息，又获得了更多浅层特征图的信息，且兼顾了大小尺度的目标。已有实验结果表明，浅层的特征图能表达图像细节信息，而深层的特征图能表达图像语义信息，因此仅利用深层的语义信息而忽略浅层信息，会降低模型的性能。相比单尺度检测网络，3 个尺度检测层边框回归精度更高，对小目标的识别效果更佳，网络能够获得更多鲁棒的图像信息。

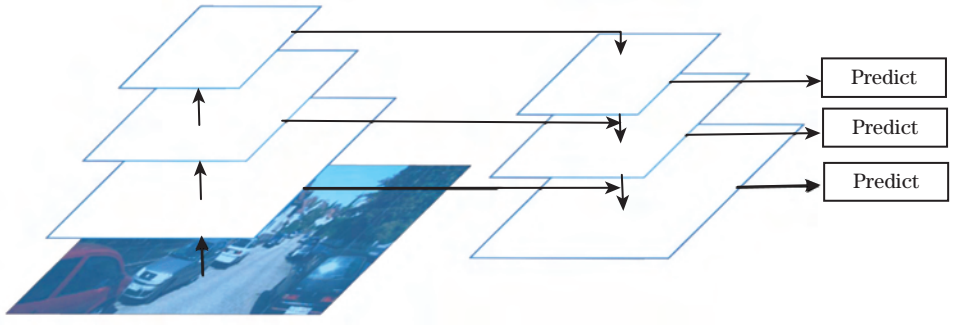

图 2　特征金字塔网络

Fig. 2　Feature pyramid network

改进的 YOLOv3 多尺度网络结构如图 3 所示。在 YOLOv3 模型的基础上，将 3 个检测层扩展为 4 个，4 个检测层尺度大小分别为 13×13、26×26、52×52、104×104。对多个卷积层进行卷积操作后，对底层 13×13 的特征层进行 2 倍上采样变成 26×26 大小，使其与上一层的尺度大小一致；之后与 26×26 特征层进行拼接，并将融合后的特征图输入到上一个检测层中，直到实现 4 个检测尺度的融合；最后对每个融合后的特征层进行单独预测。这 4 个检测尺度都具有高分辨率和高语义信息，且没有增加网络复杂度，可以更好地适应小目标的检测。

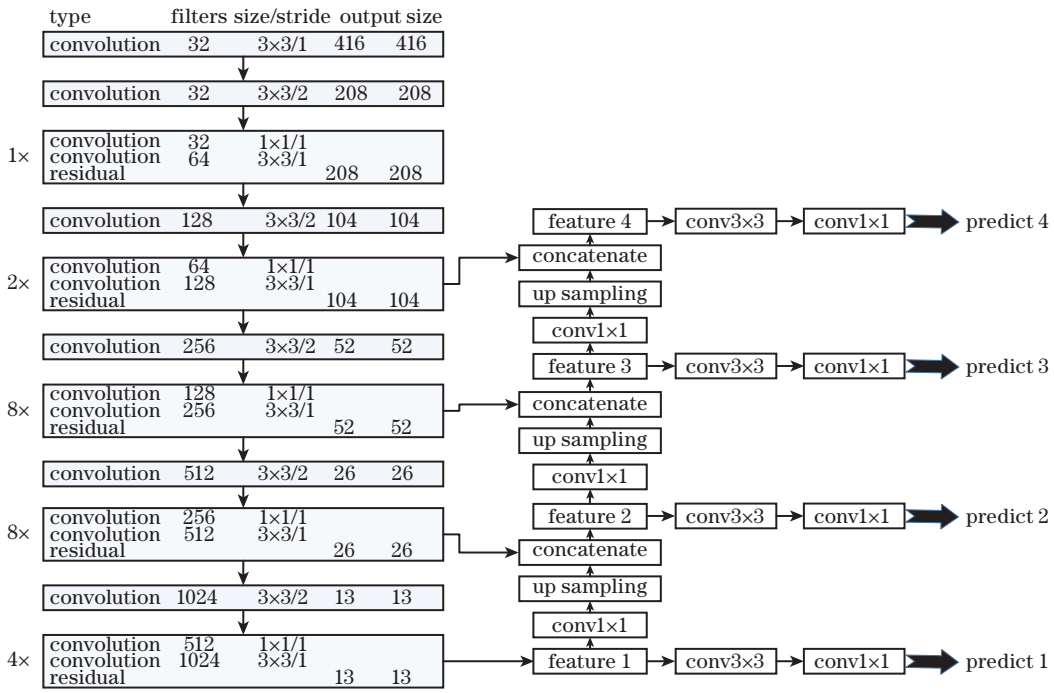

图 3　改进的多尺度 YOLOv3 结构

Fig. 3　Improved multi-scale YOLOv3 structure

在深度神经网络中，为了增加感受野并降低计算量，要进行池化等操作。池化可以减少参数并扩大感受野，但会降低空间分辨率。为了在不丢失分辨率的同时扩大感受野，可以使用空洞卷积。在传

统卷积的基础上，空洞卷积通过在卷积核中注入空洞来扩大感受野。图 4(a)、(b)分别为空洞卷积运算和传统卷积运算。可以明显看出，空洞卷积能够扩大卷积核的感受野，更好地提取上下文多尺度特征，这在视觉检测任务中非常重要。改进的 YOLOv3 模型在每次 2 倍下采样时加入一个超参数膨胀率(dilation rate)，利用膨胀后的卷积核替代移动步长为 2 的卷积来实现对图像的下采样操作。通过将传统卷积改为空洞卷积，可以在保持参数量不变的情况下，靠后的卷积层也可保持较大的感受野尺寸，更好地保留浅层特征，更有利于对小目标进行检测，提高模型整体性能。

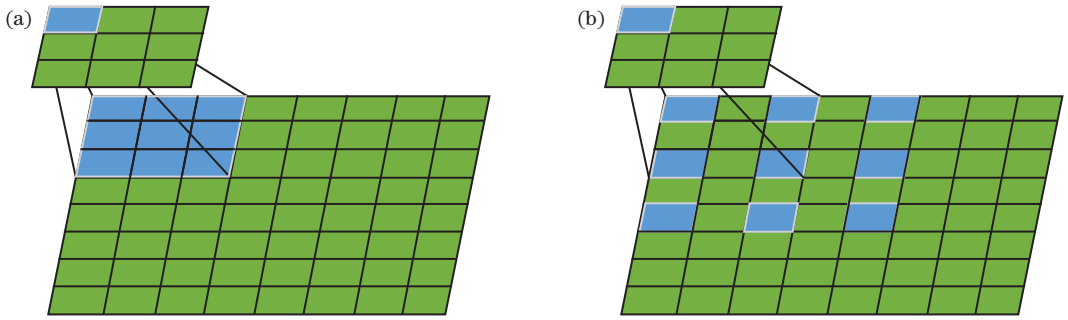

图 4　空洞卷积和传统卷积。(a)空洞卷积；(b)传统卷积

Fig. 4　Dilated convolution and traditional convolution. (a) Dilated convolution; (b) traditional convolution

检测任务的损失函数一般由分类损失函数(Classificition Loss)和回归损失函数(Bounding Box Regeression Loss)两部分组成。YOLOv3 在目标框坐标回归过程中采用的是均方误差(MSE)损失函数，但以 MSE 作为边框坐标回归的损失函数有两个明显的缺点：MSE 损失值低并不等价于交并比(IoU)高；MSE 损失对目标框尺度变化比较敏感，不具有尺度不变性。

IoU 作为目标检测算法常用的评价标准，是预测框和真实框的交集与并集之比，计算公式为

$$R_{\mathrm{IoU}}=\frac{|\boldsymbol{M}\cap\boldsymbol{N}|}{|\boldsymbol{M}\cup\boldsymbol{N}|},\tag{1}$$

式中：$\boldsymbol{M}=(x,y,w,h)$表示预测框；$\boldsymbol{N}=(x^{\mathrm{gt}},y^{\mathrm{gt}},w^{\mathrm{gt}},h^{\mathrm{gt}})$表示真实框；$x$、$y$ 分别表示边框中心点的横、纵坐标，w、h 分别表示边框的宽、高。虽然 IoU 可以直观反映预测框和真实框的关系，但也仅在两个框发生重叠时有效，IoU 对于非重叠部分不会提取任何有效信息。针对 IoU 作为评价标准的缺陷，文献[19]使用 CIoU 来降低 IoU 对检测精度的影响，将 CIoU 损失运用于 SDD 和 Faster RCNN 上，并进行了实验对比。实验结果如表 1 所示，其中 AP75 表示 IoU 阈值为 0.75 时的准确度，CIoU 损失将两个模型的 mAP 分别提高了 1.90%和 0.84%。而 AP75 则分别提升了 0.78%和 2.59%。

表 1　CIoU 在 SSD 和 Faster RCNN 上的实验结果对比

Table 1　Comparison of experimental results of CIoU on SSD and Faster RCNN

Methods	Relative improvement/%	
	mAP	AP75
Faster RCNN	1.90	0.78
SSD	0.84	2.59

针对 MSE 损失函数和 IoU 损失函数的缺陷，将 CIoU 损失作为目标框回归的损失函数。与 IoU 近似，CIoU 也是一种距离度量标准。其公式为

$$R_{\mathrm{CIoU}}=R_{\mathrm{IoU}}-\frac{\rho^2(m,n)}{d^2}-\upsilon\beta,\tag{2}$$

式中：m 和 n 分别表示 $\boldsymbol{M}$ 和 $\boldsymbol{N}$ 的中心点；$\rho^2(m,n)$表示预测框和真实框的中心点欧氏距离；d 表示能够同时包含预测框和真实框的最小闭包区域的对角线距离；υ 表示权衡参数；β 衡量了长宽比的一致性。υ 和 β 的公式为

$$\upsilon=\frac{\beta}{1-R_{\mathrm{IoU}}+\beta},\tag{3}$$

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

CIoU 的相应损失函数为

$$L_{\mathrm{CIoU}}=1-R_{\mathrm{IoU}}+\frac{\rho^2(m,n)}{d^2}+\upsilon\beta。\tag{5}$$

CIoU 将真实框与预测框的中心距离、重叠率及长宽比作为参考，对预测框的回归更加稳定，不会像 IoU 损失函数一样出现训练发散现象。CIoU 对预测框尺度变化并不敏感，即具有尺度不变性。

2.2　实验细节及结果

实验环境：操作系统为 Windows10；计算机 CPU 为 i7-9750H；内存为 16 GB；GPU 为 RTX2080s，其显存为 8 GB；实验所用框架为 PyTorch1.2.0+CUDA10.0；实验中使用 TorchVision 库的 transforms，datasets 工具对图像数据进行归一化预处理。

实验数据集选用 KITTI 数据集，KITTI 数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集。使用其中的 2D 目标检测数据集，并对数据集的标签进行修改，保留了实验所需的 3 个类别，Car、Van、Truck，如图 5 所示。其中训练集和测试集分别有 7481 张和 7518 张图像。选取其中 9000 图像作为实验数据，数据集经上下翻转、水平翻转以及随机裁剪后扩充至 16386 张图像。并按照 9∶1，从该数据集中随机选取 14747 张作为训练数据集，剩余的 1639 张图像作为测试数据集。

图 5　三种车型实例。(a)货车；(b)面包车；(c)小轿车

Fig. 5　Three examples of vehicles. (a) Truck; (b) Van; (c) Car

实验在训练阶段进行了 15000 次迭代，图 6 为训练次数与损失的关系图，每张输入图像尺寸都被固定至 416 pixel×416 pixel。总体参数设置参照 YOLOv3 模型，将 batch_size(网络一次训练所需要的数据量，其大小影响网络的收敛速度)设置为 64，优化函数中动量参数与权重衰减系数，分别设置为 0.9 和 0.0005。训练中出现损失率下降速度过于缓慢以及难以收敛等现象，推断原因为 batch_size 设置过大。为了加快网络的训练速度并解决难以收敛的问题，分别采用 batch_size 的值为 8，16，32 进行训练。结果表明，当 batch_size 为 8 时，网络收敛速度最快，因此将 batch_size 设置为 8。学习率初始为 0.001，但损失大范围下降 4000 epochs 后出现了损失率升高和变化极小的趋势。为了让损失率继续降低以达到更优的拟合效果，将学习率降低为 0.0001。此条件下损失值小范围浮动，模型可以更好地进行训练，最终损失值在 0.78，基本保持不变。

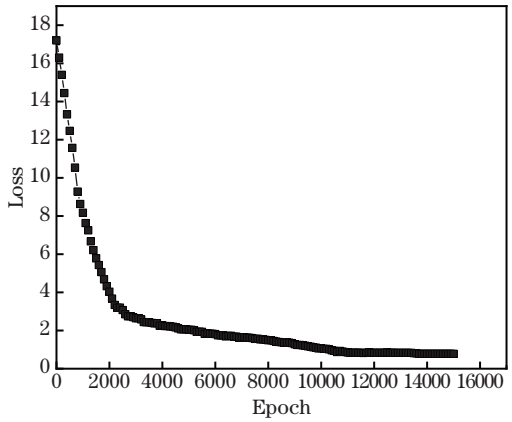

图 6　训练次数与损失值的关系

Fig. 6　Relationship between training times and loss value

3　实验结果分析与讨论

3.1　评测指标对比与分析

首先，利用模型对某类目标检测的准确率(P)和召回率(R)计算出某类别的 AP；然后得到作为本次检测模型性能评估标准的 mAP 和每秒传输帧(FPS)。公式分别为

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}},\tag{6}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}},\tag{7}$$

$$P_{AP}=\frac{1}{N_c}\sum_{r_c\in R_c}p(r_c),\tag{8}$$

$$P_{mAP}=\frac{\sum P_{AP}}{N(c)},\tag{9}$$

式中：N_{TP} 表示检测框中正确，且 IoU 大于阈值的实例；N_{FP} 表示检测框中错误，且 IoU 小于阈值的实例；N_{FN} 表示真正正确的框，却没有被模型检测到的实例；N_c 表示划分为第 C 类 P 和 R 的数量；$p(r_c)$表示在 c 类召回率为 r_c 时的 p 值；$N(c)$表示类别数。在实验中，设置 IoU 来检测预测框和真实框交集与并集之间的比例，IoU 阈值设为 0.5。

对 YOLOv3、YOLOv2、SSD512 和 Faster RCNN 模型进行了重现，并与所提模型的 mAP 和 FPS 进行了比较，性能对比如表 2 所示。均采用 KITTI 数据集对这些算法进行训练和测试。从表 2 能够看出：所提改进 YOLOv3 算法的 mAP 为 95.78%，并且检测速度达到了 44.1 frame/s；改进

的 YOLOv3 算法的精度比 Faster RCNN 高出 18.88 个百分点；经典的目标检测算法 Faster RCNN 有 76.90%的 mAP，但无法做到实时检测，这是由于网络中的 RPN 结构增加了计算量；YOLOv2 检测速度最快，但精度比改进的 YOLOv3 算法低 26.30 个百分点；SSD512 在精度和检测速度上均有良好表现，但与改进的 YOLOv3 算法相比仍有较大差距；在检测速度上，所提算法比 YOLOv3 慢 4.5 frame/s，这是由于多加的尺度检测层和改进的损失函数增加了计算复杂度，但检测精度提高了 4.2 个百分点。

表 2　各算法的性能对比

Table 2　Performance comparison of various algorithms

Algorithm	Input size	FPS /(frame·s^{-1})	mAP /%
Faster RCNN	600×600	11.63	76.90
YOLOv2	416×416	114.26	69.48
YOLOv3	416×416	48.60	91.58
SSD512	512×512	27.69	79.70
Improved YOLOv3	416×416	44.10	95.78

3.2　模型效果测试和对比分析

所提算法在 KITTI 测试集进行了预测。在预测结果中抽取 3 张图像后与 YOLOv3 检测后的结果进行对比，对比结果如图 7 所示。图 7(a)、(c)、(e)为改进的 YOLOv3 对测试集进行检测后的效果图；图 7(b)、(d)、(f)为 YOLOv3 的检测效果图。图像检测结果包含车辆目标的分类准确率和边框回归效果。从图 7 的对比分析可知：YOLOv3 虽然能检测出大部分的车辆目标，但对极小目标无法处理；改进的 YOLOv3 模型将检测尺度增加至 4 个后，能够更好地检测极小的目标，对遮挡范围较大的目标也能识别，且准确率不低。这是因为边框回归损失函数改为 CIoU 后，相比 MSE 损失函数，对遮挡的目标有更好的置信度。

图 7　改进的 YOLOv3 与原始 YOLOv3 的检测对比。(a)(c)(e)改进的 YOLOv3；(b)(d)(f) YOLOv3

Fig. 7　Detection comparison of improved YOLOv3 and original YOLOv3.
(a)(c)(e) Improved YOLOv3; (b)(d)(f) YOLOv3

3.3　空洞卷积对比分析

本文捕获部分下采样层特征图，以验证空洞卷积效果。图 8(a)为加入空洞卷积后截取下采样层的图片，图 8(b)为采用传统卷积的下采样层图片。可以明显看出，在下采样层加入空洞卷积后，改进的 YOLOv3 模型将图像纹理和细节保存得更好，更有益于网络对小目标的检测。

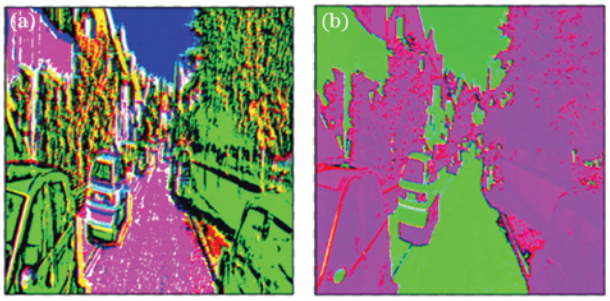

图 8 空洞卷积与传统卷积效果对比。
(a)空洞卷积;(b)传统卷积

Fig. 8 Comparison of dilated convolution and traditional convolution. (a) Dilated convolution; (b) traditional convolution

4 结 论

使用改进后的 YOLOv3 模型对 KITTI 数据集中的车辆目标进行检测,对数据集使用归一化预处理方式,让数据的分布集中在激活函数梯度较大的区域;将 YOLOv3 的 3 个检测尺度改为 4 个,显著地增强了模型对小目标的检测效果,尤其是极小目标的检测;在下采样层将传统卷积换为空洞卷积,使网络中采样后特征图的分辨率更高,这保存了更多的图像细节且扩大了感受野,并为网络的训练提供了充分的上下文信息;将损失函数改为 CIoU 后,目标框位置的回归更加准确,且模型对遮挡物体的识别置信度更高,检测效果更好。实验结果表明,改进的 YOLOv3 在精度和边框位置回归准确度上都有一定程度的提升。但该模型在检测速度上仍有不足,下一步需要在保证准确度的基础上,进一步提高检测速度。

参 考 文 献

[1] Zhang F K, Yang F, Li C. Fast vehicle detection method based on improved YOLOv3[J]. Computer Engineering and Applications, 2019, 55(2): 12-20.
张富凯, 杨峰, 李策. 基于改进 YOLOv3 的快速车辆检测方法[J]. 计算机工程与应用, 2019, 55(2): 12-20.

[2] Li H, Fu K, Yan M L, et al. Vehicle detection in remote sensing images using denoizing-based convolutional neural networks[J]. Remote Sensing Letters, 2017, 8(3): 262-270.

[3] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 3354-3361.

[4] Yu F, Chen H F, Wang X, et al. BDD100K: a diverse driving dataset for heterogeneous multitask learning [C] // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 2633-2642.

[5] Brostow G J, Fauqueur J, Cipolla R. Semantic object classes in video: a high-definition ground truth database[J]. Pattern Recognition Letters, 2009, 30(2): 88-97.

[6] Zhuo D, Jing J F, Zhang H H, et al. Classification of chopped strand mat defects based on convolutional neural network [J]. Laser & Optoelectronics Progress, 2019, 56(10): 101009.
卓东, 景军锋, 张缓缓, 等. 基于卷积神经网络的短切毡缺陷分类[J]. 激光与光电子学进展, 2019, 56(10): 101009.

[7] Yuan L S, Lou M Y, Liu Y Q, et al. Palm vein classification based on deep neural network and random forest [J]. Laser & Optoelectronics Progress, 2019, 56(10): 101010.
袁丽莎, 娄梦莹, 刘娅琴, 等. 结合深度神经网络和随机森林的手掌静脉分类[J]. 激光与光电子学进展, 2019, 56(10): 101010.

[8] Zhao H, An W S. Image salient object detection combined with deep learning [J]. Laser & Optoelectronics Progress, 2018, 55(12): 121003.
赵恒, 安维胜. 结合深度学习的图像显著目标检测[J]. 激光与光电子学进展, 2018, 55(12): 121003.

[9] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[10] He K M, Zhang X Y, Ren S Q, 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.

[11] Girshick R. Fast R-CNN [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[12] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[13] Dai J F, Li Y, He K M, et al. R-FCN: object

detection via region-based fully convolutional networks [C] // 2016 Conference on Neural Information Processing Systems, December 5, 2016, Red Hook, NY, United States. New York: Curran Associates, 2016: 379-387.

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[15] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[16] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[17] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2019-11-01]. https://arxiv.org/abs/1804.02767.

[18] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[19] Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.