文章编号:1671-4598(2023)04-0232-07 DOI:10.16526/j.cnki.11—4762/tp.2023.04.036 中图分类号:TP391.41;TP183 文献标识码:A

基于 Mask Scoring R-CNN 的高质量数据集快速自动标定方法

胡馨月[1],谢 非[1,3],王 军[2],马 磊[2],黄懿涵[1],刘益剑[1,3]
(1.南京师范大学 电气与自动化工程学院,南京 210023;2.南京三万物联网科技有限公司,南京 210000;
3.江苏省三维打印装备与制造重点实验室,南京 210042)

摘要:针对计算机视觉领域人工标定多目标数据集时间冗长的问题,提出一种基于 Mask Scoring R-CNN 的高质量数据集快速自动标定方法;首先,设计了高质量数据集快速自动标定架构,训练数据自动标定模型并搭建目标分类与标定系统;其次,在对比不同残差网络及引入迁移学习基础上,进一步研究了基于 MaskIoU Head 的多目标掩膜标定质量评价方法,完成基于 Mask Scoring R-CNN 的多目标高质量数据集快速自动标定方法设计;最后,以车辆数据为例进行数据集快速自动标定方法验证,实验结果表明,相较于 Mask R-CNN 和 Faster R-CNN 方法,Mask Scoring R-CNN 方法具有目标数据分类效果好及掩膜分割精度高的优点,检测准确率达到 93.4%,且标定速度相较于人工标定速度提升了 95.77%。

关键词:目标检测;实例分割;迁移学习;高质量数据集;快速自动标定

Rapid Automatic Labeling Method for High Quality Data Sets Based on Mask Scoring R-CNN

HU Xinyue[1], XIE Fei[1,3], WANG Jun[2], MA Lei[2], HUANG Yihan[1], LIU Yijian[1,3]
(1. College of Electrical & Automation Engineering, Nanjing Normal University, Nanjing 210023, China;
2. College of Automation & Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
3. Jiangsu Province 3D Printing Equipment and Manufacturing Key Lab, Nanjing 210042, China)

Abstract: Aiming at the problem of lengthy manual calibration of multi-target data sets in the field of computer vision, a rapid automatic labeling method for high quality data sets based on Mask Scoring R-CNN is proposed. Firstly, the rapid automatic labeling framework for high quality data sets is designed. Then, the automatic labeling model of multiple target data is trained, and the classification and labeling system for the target is built. Secondly, on the basis of the comparison of different residual networks and the introduction of transfer learning, the quality evaluation method of multiple targets mask labeling based on MaskIoU Head is further studied. Besides, the rapid automatic label method for multiple targets high quality data sets based on Mask Scoring R-CNN is implemented. Finally, the vehicle data is taken as an example, the data sets rapid automatic labeling method is verified. The experimental results show that compared with the Mask R-CNN and Faster R-CNN, the Mask Scoring R-CNN has the advantages of good effect of target data classification and high accuracy of mask segmentation, its detection accuracy reaches 93.4%, and the labeling speed of the method is 95.77% higher than that of manual labeling.

Keywords: target detection; instance segmentation; transfer learning; high quality data sets; rapid automatic labeling

0 引言

无论是计算机视觉领域或是深度学习领域,大量的数据集必不可少。目前,已经存在大量被广泛应用的数据集例如 coco、Labelme 等,然而仍有大部分特定领域缺少足够的数据集,例如在智慧交通领域车型识别和检测[1],车辆数据集的分类与标定对于研究车辆的各类特征具有重大意义,例如对车辆的目标检测、车辆分类、车牌识别、车辆测速和车色识别等[2-5]。目前与车辆相关的数据集有:KITTI、UA-DETRAC BDD100K 数据集等[3-6]。但是这些数据集中多为正向的车辆,并不适合全部实际交通情况下的车

收稿日期:2022-11-27; **修回日期**:2022-12-28。
基金项目:国家自然科学基金项目(41974033);江苏省科技成果转化(BA2020004);江苏省省级工业和信息产业转型升级专项资金项目。
作者简介:胡馨月(2002-),女,江苏南通人,大学本科,主要从事机器视觉与目标检测方向的研究。
通讯作者:谢 非(1983-),男,江苏徐州人,博士,副教授,硕士生导师,主要从事机器视觉与图像处理,机器学习与深度学习方向的研究。
引用格式:胡馨月,谢 非,王 军,等.基于 Mask Scoring R-CNN 的高质量数据集快速自动标定方法[J].计算机测量与控制,2023,31(4):232-238.

辆识别任务。因此针对特定领域的数据集制作非常重要。通过人工方式对多目标数据集进行标注，不仅耗时耗力，并且疲劳状态下标记的数据质量较低。这种方法难以快捷方便的获取质量高、数量多且满足要求的多目标数据集。

本文结合 Mask Scoring R-CNN 网络框架与迁移学习和深度残差网络，并建立多目标数据质量评分机制，并且以车辆数据集为例，通过基于 Mask Scoring R-CNN 的高质量数据集快速自动标定方法，对遮挡、目标小、种类多和环境复杂情况下的车辆目标进行实例分割并生成对应标签文件，最终得到高质量车辆标定数据集。

随着深度学习方法的广泛应用，研究人员针对车辆目标检测方法和实例分割的不足也在不断拓展研究[7-17]。彭博等通过改进的 Faster R-CNN 对道路中车辆进行分类和识别[7]；陈辰等通过级联 Adaboost 算法针对各个子问题分别训练检测模型，提高车辆目标检测精度[8]；袁功霖等人利用迁移学习和图像增强，使得小规模数据即可训练出有效的识别网络[9]。Yebes 等采用两阶段目标检测网络实现城市道路中的车辆检测与目标分类[10-14]，但是均侧重于车辆检测中分类精度、掩膜标定精度或训练时间其中某一方面，没有对三方面进行综合考虑。Kim 等采用轻量级神经网络进行车辆检测，提高了车辆检测的实时性[15-16]。以上方法虽然可以实现图像的检测与定位，但是在遮挡的环境下目标图像分割精度较低。不适用于实时交通下的车辆识别任务及车辆数据集制作。

本文针对人工标定多目标数据集时间冗长，训练实例分割模型需要大量数据和较长训练时间，且传统实例分割算法中评价目标掩膜分割质量方法不准确的问题，开展基于 Mask Scoring R-CNN 的高质量数据集快速标定方法研究，该方法可以自动对大量无标签数据进行自动标注，生成大量的车辆实例分割图像。然后，为了筛选出高质量的标签文件，提出了基于 MaskIoU Head 的质量判别方法，并以此建立了网络评分机制，筛选出高质量的数据集。本文方法具有以下优点：

1）传统多目标实例分割方法仅仅针对识别分类精度、识别速度某一方面开展研究，本文结合迁移学习、3 种深度残差神经网络和优化网络中各项超参数有效的提高了实例分割精度并大幅度降低训练时间和减少训练样本，为后续的数据集标定奠定了基础。

2）Mask R-CNN 方法采用掩膜重叠像素点的方法衡量掩膜质量，但是掩膜是不规则图像。这种方法并不准确。本文结合 Mask Scoring R-CNN 中的 MaskIoU Head 分支，建立多目标标定图像评分机制，对网络进行监督训练，可以在遮挡、目标小、种类多和环境复杂的情况下提高实例分割精度，并对掩膜标定质量进行准确衡量[17]。在相同数量的图像输入下，相较于需要四小时左右的人工标定，本文方法仅需 7 分 56 秒，实现了大量高质量数据集快速标定。

1　高质量数据集快速自动标定架构

本文以高质量车辆数据集标定方法为例，总体框架如图 1 所示，下面针对此方法的两部分分别展开说明。

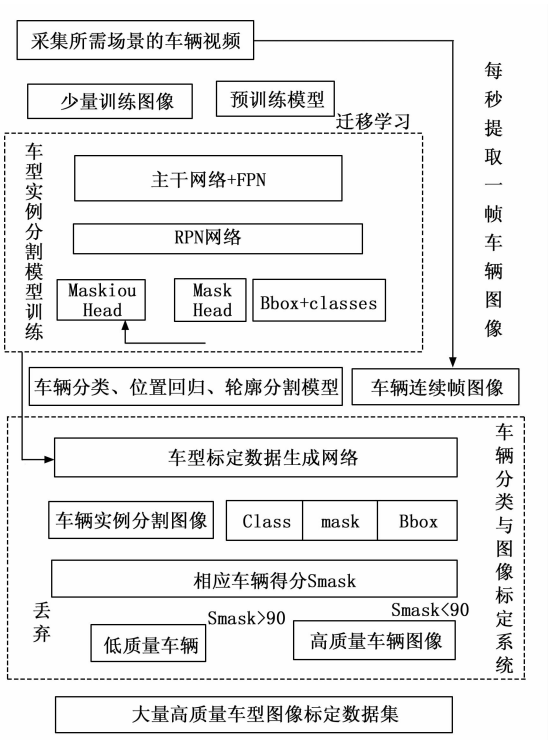


图 1　高质量数据集自动标定方法

1）车辆数据自动标定模型训练部分：利用采集的车辆视频每秒提取一帧图像，取出少量图像进行人工标定（约每个类别 80 张图像），将这些图像输入到车型实例分割网络中，结合 coco 数据集的 80 分类预训练模型进行迁移学习，减少训练时间和防止网络过拟合。得到车型分类、位置回归和轮廓标定模型[18]。

2）搭建车辆分类与标定系统：搭建基于 Mask Scoring R-CNN 的车辆分类与图像标定系统，采集与第一部分类似场景下较长的一段车辆视频，输入车辆分类与标定系统，每秒提取两帧图像，输入到第一部分得到的车型分类、位置回归和轮廓标定模型中，得到每幅图中的车辆类别、边界框和车辆掩膜，然后，得到标定后车辆图像中每辆车的分数，即 Smask，如果一幅图像中全部车辆的 Smask 都大于 90，则这幅图像为高质量车辆图像，反之为低质量图像，保留高质量车辆图像与对应的标签文件，生成车辆高质量数据集。

2　基于 Mask Scoring R-CNN 的高质量数据集快速标定算法

2.1　网络结构

本文提出的高质量数据自动标定方法是基于 Mask Scoring R-CNN 网络框架，标定网络如图 2 所示。包括主干网络（Backbone network）、图像金字塔网络（FPN，feature pyramid networks）、区域建议网络（RPN，region proposal network）、分类与回归分支（R-CNN Head）、掩膜分支（mask head）以及掩膜评分分支（MaskIoU head）组成。

此网络不仅能输出具体类别和目标框，还能对物体目标轮廓进行精准分类和标定。

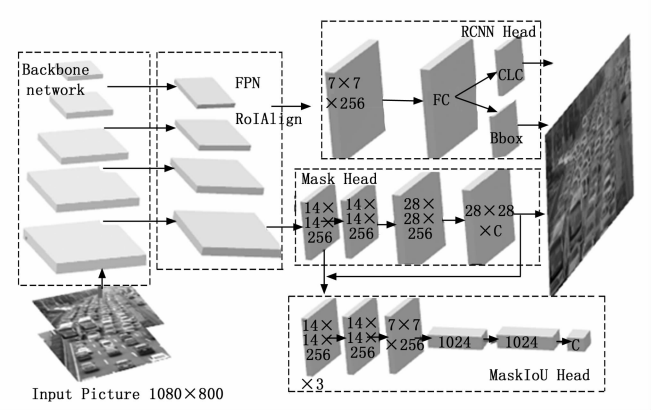


图 2 车辆图像标定网络

本文车辆数据自动标定网络可分为 4 个部分：第一部分为车辆图像特征提取，先通过主干网络（backbone network）提取图片特征，再通过 FPN 网络形成图像特征金字塔。第二部分为区域生成网络（RPN），该部分使用 RoIAlign 从每个候选区域（proposal）提取特征，筛选出目标车辆所在区域。第三部分通过 R-CNN Head 对候选区域进行目标区域分类、边界框的回归，同时通过 Mask Head 将车辆从复杂环境分离出来并对其轮廓进行预测、标定。第四部分为车辆掩膜质量评分部分。本文利用 MaskIoU Head 建立标定的车辆图像评分机制，对第三部分中车辆掩膜标定质量进行打分，通过分数衡量预测车辆掩膜与真实车辆区域的一致性，同时对车辆图像进行取舍。经历 4 个部分后，得到车辆图像标定模型[19]。

2.2 网络组成部分具体设计

2.2.1 基于残差网络的多目标特征提取

通过主干网络从无标签的车辆图像中提取特征，并通过 FPN（特征金字塔）形成多尺度的特征层，增强网络对小目标的识别能力。

主干网络：用来特征提取的 CNN 网络，主要检测图像中的高级特征，其中，主干网络可以是任意的卷积层进行组合构成的特征提取网络，或者是常用的高精度卷积神经网络（如：ResNet 50、ResNet101、VGG19 等）。利用主干网络，通过卷积操作将图像从 1980 * 1080 * 3（RGB）转变为 32×32×2048 的特征图。这个特征图将作为特征金字塔网络的输入[20]。

特征金字塔网络：使得 feature map 包含的特征更全面。此特征金字塔一共有五层，从第一个层提取特征后逐层传递到第五层，但尺度逐层下降一倍，生成不同尺度的 feature maps，再将相邻 feature maps 相减，得到新的 feature map。使得新的特征图既保留了低层次中包含原图更多信息的特点，又包含高层次特征图像中更深层次特征。本文选择第四层特征图作为后续网络的输入。

2.2.2 迁移学习和生成目标候选区域

迁移学习：迁移学习是给网络中的权值一个初始值，coco 数据集的 80 分类预训练模型与本文需要训练的车辆数据标定模型均为图像识别模型，可以有效防止过拟合与减少训练数据量，降低训练时间。且 coco 数据集中图像拍摄于城市道路，含有非车辆图像的类别，可以提高车辆图像的背景与前景分类精度。因此引入迁移学习。

建议区域网络（RPN）：用于生成建议区域（region proposals）作用于特征金字塔提取的 feature map 中，利用滑动窗口在 feature map 中进行扫描，找到包含目标的区域，RPN 扫描过的区域称为锚点，锚点越多精度越高，相应训练速度会降低。为了在精度与速度之间保持平衡，本文的实验中每张图像大约有 10 万个不同大小和高宽比的锚点，以此覆盖图像中更多的面积，提高检测精度。

RoI 分类器：作用于 RPN 网络产生的建议区域中，可将属于背景还是目标的区域进行分类，属于目标物体的建议区域称为正区域，属于背景的建议区域称为负区域。保留正区域，丢弃负区域。

2.2.3 目标位置回归、分类和轮廓提取

R-CNN Head：位置回归与目标分类。通过此分支将目标分类，在 feature map 上对边界框进行回归。同时对第二部分的正区域进行合并，并判别目标的类别。采用 IoU 方法对预测的边界框进行评估。

边界框 IoU 如图 3 所示。虚线框为目标的真实边界框（Ground truth），黑色填充部分为 R-CNN Head 预测的 Bbox（边界回归框），图 3 中从左到右 Bbox 与 Ground truth 之间重叠越多，说明此网络边界框预测效果越好，如图 3 最右侧图形所示。

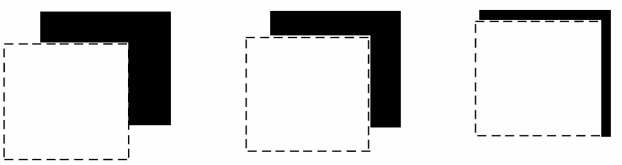


图 3 边界框 IoU

Mask Head：由全卷积神经网络（FCN，fully convolutional networks）构成，在 RoI 分类器筛选后的正区域上生成目标的掩膜，这层掩膜可以准确地包围目标物体，再通过反卷积放大到原图，得到目标图像的轮廓，并将每个图像中目标轮廓上的像素点坐标保存，生成对应的标签文件，也是多目标数据集标定的关键之一。

但是，主干网络中进行的卷积操作会导致原图信息有所丢失，在 Mask Head 中，将 feature map 反卷积到原图后会出现预测掩膜与真实掩膜有一定偏差。因此需要一种方法去衡量预测掩膜的质量。在传统方法 Mask R-CNN 中用二者交叉面积与二者累加面积的比值方式计算 MaskIoU，来衡量预测掩膜质量，但是需要保证二者有相同的高和宽。可是这种方法计算的 MaskIoU 与预测掩膜并不为线性关系，因此这种方法是不准确的。

2.2.4 基于 MaskIoU Head 的多目标掩膜标定质量评价

MaskIoU Head：利用卷积神经网络中回归原理，精准

地评定目标的 mask 质量，并在网络训练中进行监督，很好地解决了 Mask R-CNN 对目标 mask 质量评分不准的问题。卷积神经网络常常用来回归两个相似图像，本文利用这个卷积神经网络分支对真实掩膜（Truth-mask）与预测掩膜（predict-mask）进行回归，并计算出每个目标 mask 的 MaskIoU 值，得到的 MaskIoU 值为 S_{IoU}，也是对每个目标蒙版质量的评价分数。该质量评价方法用于评定标签文件的质量，评估的内容主要包括包围目标轮廓的精度和目标分类精度。然后通过设置质量阈值，将质量低于阈值的标签丢弃，保留质量高于阈值的标签。最后，将高于阈值的标签和对应的车辆图像数据共同构成车辆语义分割数据集，这也是该网络的数据增强结果。

MaskIoU Head 输入结构：本文将 Truth-mask 和 predict-mask 一起作为 Mask Head 的输入。其中 Truth-mask 存在于 RoI feature 中，predict-mask 为 Mask Head 输出的目标预测掩膜。由于 predict-mask 与 RoI feature 尺寸不同，因此设计了两种输入结构。MaskIoU Head 的两种输入结构如图 4 所示。

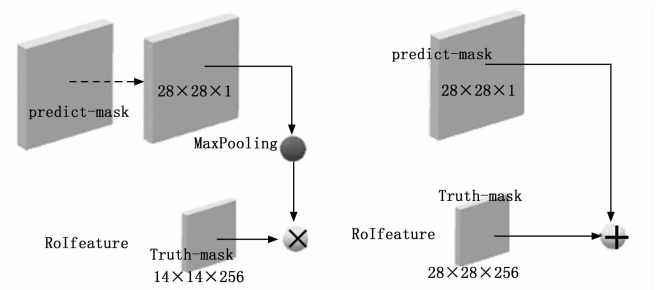


图 4　MaskIoU Head 的输入结构

具体说明如下：图 4 中左图设计的输入结构是将所有的 mask 经过 kernel size 为 2，stride 为 2 的 max pooing，然后与 RoI 输出的 RoI feature 相乘。右图设计的输入结构为目标 mask 不经过最大池化直接与高分辨率的 RoI feature 相加。两种结构均可作为 Mask Head 的输入。

MaskIoU Head 网络结构设计：由 4 个卷积层和 3 个全连接层组成。对于 4 个卷积层，本文将所有卷积层的核大小和滤波器个数分别设置为 3 和 256。对于 3 个全连接层，本文结合 R-CNN Head 设计原理，将前两个 FC 层输出设置为 1 024 以连接所有神经元，最后一个 FC 层的 C 为需要分类的类别数，输出属于不同类别的蒙版分数 S_{IoU}. MaskIoU Head 结构图如 5 所示。

2.3　网络损失函数与评分机制

设计完网络 4 个部分后，需要通过损失函数来度量网络的性能，以及设计评分机制来评价目标分割效果。

网络损失函数设计：本文网络结构主要由 R-CNN Head、RPN、Mask Head 和 MaskIoU Head 等各个分支组成，因此本文损失函数公式为：

$$L = L_{class} + L_{bbox} + L_{p} + L_{r} + L_{iou} \tag{1}$$

其中：L_{class}为目标检测分类的损失，L_{bbox}是回归目标检测框的损失，L_{mask}为目标 mask 分割的损失，L_{p} 为 RPN 网络损失，L_{r} 为权重正则化损失。L_{IoU}为 MaskIoU Head 损失函数。

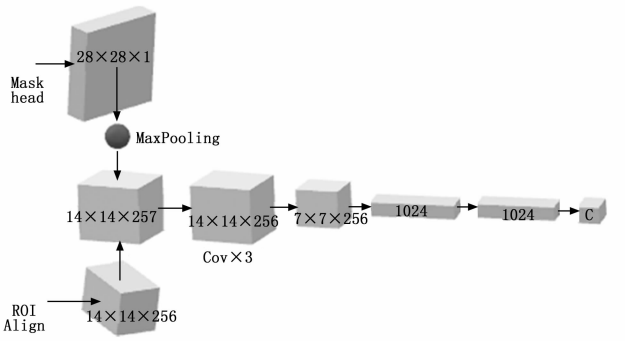


图 5　MaskIoU Head 结构图

网络评分机制：本文目标车辆只属于一个类别，这就要求本文方法在两方面必须同时表现良好：1）需要对不同车型进行精确的分类；2）预测的车辆 Mask 和真实车辆 Mask 之间需要有较高的重合度，用一个目标函数来表达这两个任务较为困难。因此，本文将 mask 质量评判标准分解成目标分类和掩膜回归评分，公式为：

$$S_{mask} = S_{cls} \times S_{IoU} \tag{2}$$

其中：S_{mask}为评定目标检测质量的分数，S_{cls}为 RNN Head 中对目标分类效果评定的分数，S_{IoU}表示 predict-mask 与 Truth-mask 之间重合程度的分数。如果一张图像中所有目标的 Smask 均高于 90 分，那么这张图像即为高质量的目标图像，将这些图像及对应标签文件存储作为相关数据集。

3　实验与分析

3.1　数据预处理及少量训练数据标注

为了自动生成大量车辆高质量数据，首先需要训练一个车辆数据标定模型。训练数据为亲自采集的车辆视频，每秒提取两帧车辆图像，取出少量车辆图像，使用 Labelme 软件进行人工标注。

其中 Labelme 标记六种车型如图 6 所示，图像数据中含有 800 张含多种车辆的图像。并将其划分为 680 张训练图像，120 张验证图像。本文设计了 6 种车型进行实验：Bus（巴士）、Car（小轿车）、MircoBus（面包车）、SUV（运动型多用途汽车）、Truck（卡车）、SprotsCar（跑车）。

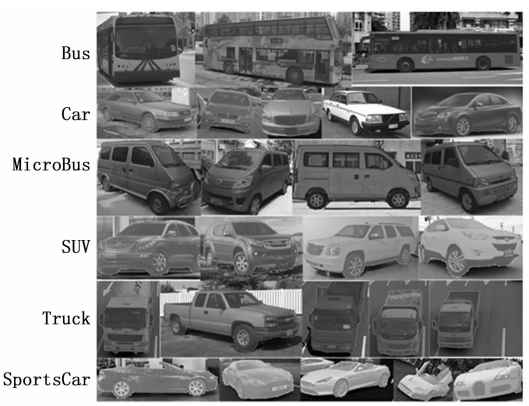


图 6　Labelme 标记六种车型

3.2 网络参数与主干网络以及迁移学习选择

网络中不同参数对神经网络训练的结果影响很大，降低迭代次数、学习率等会导致网络性能的降低。本文网络结构与 Mask Scoring R-CNN 结构相近，因此选择与 Mask Scoring R-CNN 一致的固定训练参数。网络中固定训练参数如表 1 所示。

表 1 网络中固定训练参数

参数名称	参数数值
RPN 网络 NMS 阈值	0.8
训练 NMS 后输出窗口数量	1 000
NMS 阈值	0.3
RoI 正样本比例	0.33
基础学习率	0.005
权重衰减	0.000 5
Learning_momentum(动量)	0.9
优化方法	SGD

为了设计出最佳的车辆数据集自动标定模型，本文在同一主干网络 ResNet50 下，总共进行了共 10 组对比试验，选择合适的网络参数以及观察引入迁移学习对本方法在训练时间、准确度和训练数据量方面的影响，如表 2 所示。

表 2 超参数选择与迁移学习对照试验

变量名称	试验 1	试验 2	试验 3	试验 4	试验 5
训练目标数量	2 081	2 081	2 520	3 005	3 005
验证目标数量	537	537	632	826	826
训练图像数目	680	680	1 014	1 014	1 014
验证图像数目	120	120	180	180	180
完全迭代次数	100	200	200	400	400
最小蒙版尺寸	56 * 56	56 * 56	56 * 56	56 * 56	56 * 56
输入图像尺寸	1 024 * 800	1 024 * 800	1 024 * 800	1 024 * 800	1 920 * 1 080
比例大小	(32,64,128,256)	(32,64,128,256)	(32,64,128,256)	(32,64,128,256)	(32,64,128,256)
预训练模型	NO	NO	NO	NO	NO
mIoU	0.472	0.452	0.495	0.468	0.354
mAP(IoU>0.5)	0.569	0.586	0.495	0.565	0.195
mAP(IoU>0.7)	0.472	0.488	0.376	0.376	0.159
变量名称	试验 6	试验 7	试验 8	试验 9	试验 10
训练目标数量	3 005	3 005	3 005	1 573	1 573
验证目标数量	826	826	826	537	537
训练图像数目	1 014	1 014	1 014	480	480
验证图像数目	180	180	180	120	120
完全迭代次数	100	100	100	100	100
最小蒙版尺寸	28 * 28	28 * 28	28 * 28	28 * 28	28 * 28
输入图像尺寸	1 024 * 800	1 920 * 1 080	1 920 * 1 080	1 920 * 1 080	1 920 * 1 080
比例大小	(32,64,128,256)	(16,32,64,128)	(8,16,32,64)	(8,16,32,64)	(8,16,32,64)
预训练模型	NO	NO	NO	NO	Yes
mIoU	0.459	0.419	0.512	0.389	0.504
mAP(IoU>0.5)	0.558	0.573	0.686	0.245	0.692
mAP(IoU>0.7)	0.469	0.483	0.585	0.154	0.592

mIoU 和 mAP 为实例分割神经网络中常用的评价指标，用于评判网络模型性能高低。为了严格评估方法性能，在 IoU 分别为 0.5 和 0.7 下用 mAP 衡量实例分割效果，大于阈值是真阳性，小于阈值则为假阳性。每个实验的 mIoU 和 mAP 指标显示在表的最后三行。下面对实验内容和结果进行详细的分析。

试验 1 到试验 2 使用了同样的 NMS 阈值、基础学习率等经验参数，但是使用了不同数量的完全迭代次数。完全迭代次数的增加使得试验 1 的 mAP（IoU>0.5）值从 0.569 提高到试验 2 中的 0.586，提升效果较低，且在迭代 100 次的情况下依然容易收敛，表明本文设计的网络收敛效果良好。在试验 3 至试验 6 中使用了更多数据的图像用于训练和测试，由于完全迭代次数与之前的一样，结果显示试验 3 中准确度下降，后来在试验 4 中，通过增加完全迭代次数来改进这一点，使得 mAP（IoU>0.5）到达 0.565。在试验 5 中，本文评估了图像宽度和高度的影响，将训练图像的尺寸从 1 024×800 提高到 1 920×1 080，其余参数和试验 4 一样的情况下，算法的性能较差（mAP（IoU>0.5）= 0.185）。说明高分辨率图像在当前网络参数下，准确度较低。在试验 6 中，将最小蒙版尺寸从 56×56 缩小到 28×28，与试验 4 进行对比，网络性能得到提升。在试验 7 中，本文降低了 Anchor 的比例大小，输入图像分辨率提升到 1 920 * 1 080，将最小蒙版依然设置为 28×28，发现将高分辨率图像作为输入时，网络的性能接近于试验 6，维持稳定。在试验 8 中使用了与试验 7 一样的配置，并且进一步的降低了 Anchor 的比例大小，发现网络的性能有了较大的提升，于是将（8，16，32，64）作为网络的最佳 Anchor 比例大小。

选择了最佳完全迭代次数和图像分辨率、Anchor 的比例大小等最优超参数后，为了减少训练时间、防止网络过拟合，在试验 9 中削减了一半的训练数据量，发现网络性能大幅度降低。因此，在试验 10 中利用预训练的 COCO 数据集的 80 分类模型在试验 9 基础上进行迁移学习。发现网络性能与试验 8 几乎一致，达到较高水平，可以对车辆目标进行准确实例分割与标定，但是训练时间仅为试验 8 实验的一半。

通过 10 组对照实验，分析结果表明，训练数据量越大，图像分辨越高，掩膜越小，RPN 锚的尺度越小，网络性能越好，且 100 个完全迭代次数就足够实现收敛。同时，结合迁移学习可以大幅度减少本方法的训练数据、训练时间和提高检测精度。

主干网络对比试验：ResNet50、ResNet101、MobileNet V1 这些神经网络由残差块构成，以残差学习简化了网络架构，减少了计算开销，很好的解决了梯度消失问题。为了进一步优化网络，在识别速度和准确度之间达到一个平衡，表 3 为主干网络性能对比表，在 Test10 的网络配置参数下，分别在网络训练时间、每秒图像检测时间、网络模型大小、准确度（S>90 表示分数大于 90 的车辆为实例分割准确）4

个方面对其性能做了评估。

表 3　主干网络性能对比

主干网络	训练时间/h	标定速度 speed/FPS	模型大小/MB	准确度（$S>90$）
ResNet50	12.65	2.4	186.75	93.4%
ResNet101	20.73	1.6	268.86	93.8%
MobileNet V1	14.61	2.2	207.82	84.5%

从表格中可以看出，采用 ResNet50 作为主干网络，训练时间为 12.65 小时，时间最短；这 3 种网络的标定测试速度分别为每秒 2.4 张，每秒 1.6 张，每秒 2.2 张，采用 ResNet50 作为主干网络，标定车辆图像速度最快；在模型大小对比实验中，采用 ResNet50 作为主干网络，车辆标定模型大小最小；一张图像中车辆 Smask 均大于 90 的图像为准确图像，准确图像占全部图像的比例为准确度，采用 ResNe50、ResNet101 和 MobileNet V1 为主干网络，车辆图像标定准确度分别为 93.4%、93.8%、84.5%。

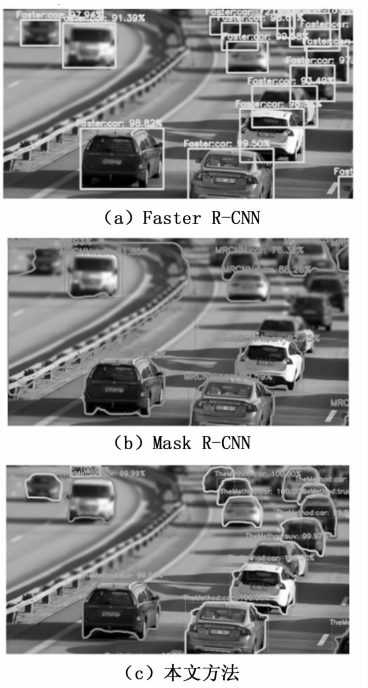
从上述实验中发现 ResNet101 精度最高、ResNet50 次之、MobileNet V1 最低。但是 ResNet101 网络层数更多，训练时间更长，ResNet50 训练时间和检测时间适中，MobileNetV1 训练时间最短，虽然 ResNet50 与 ResNet101 在车辆识别精度上准确度都很高，但是 ResNet50 在识别速度、训练时间、网络模型大小方面均优于 ResNet101。ResNet50 由于层数适中，在数据量少的情况下，既可以保证网络精度又可以防止了过拟合。更深层次的网络如 ResNet101 等需要训练数据更多的图像，反而加重了研究人员繁重的工作量。因此本方法采用 ResNet50 作为车辆标定网络中的主干网络。

经过上述试验 1 到试验 10 以及主干网络性能评估共 13 组对比试验，本文选择了合适本网络的超参数，将 ResNet50 作为主干网络，并结合了 coco 数据集的 80 分类预训练模型进行迁移学习。

3.3　多种方法的分类和掩膜分割精度性能对比

由于本文中网络是在 Mask R-CNN 基础上增加了 MaskIoU Head 来对车型图像数据标定网络进行优化，且 Faster R-CNN 是车辆检测中最常用识别网络框架。为了验证所研究方法的效果，将本文方法与传统的 Mask R-CNN 框架和 Faster R-CNN 框架在相同测试图像下进行对比实验。

车辆实例分割对比如图 7 所示。图（a）为 Faster R-CNN 目标车辆分割图像，图（b）为 Mask R-CNN 车辆分割图像，（c）图为本文算法实例分割图像．从图中可以看出，Faster R-CNN 不能对车辆轮廓进行标记。Mask R-CNN 将左边的护栏误判成了车辆，且在图像右上方有很多车辆未被识别出来，精度不够。本文算法不仅将车辆从复杂环境与重叠车辆中精确区分出来，且对于车型种类几乎没有误判，车辆轮廓标定更清晰．因此，本文方法分类准确度和实例分割精度均优于其他方法。



（a）Faster R-CNN

（b）Mask R-CNN

（c）本文方法

图 7　多种方法的车辆实例分割对比图

3.4　车辆数据自动标定速度与标定质量实验

为了验证本文方法的标定速度与标定质量，进一步进行了实验测试，采集一段的 4 分 59 秒车流视频，每秒提取一帧图像作为输入，共 358 张图像。这些车辆图像手动打标签需要 4 小时左右，本文方法标定仅需要 7 分 56 秒即可完成高质量车辆图像的筛选和标定。

输入视频截图如图 8 所示。图 9 为车辆数据标定方法的输出结果。同时为了验证本方法在车辆聚集、重叠车辆环境中分类与检测效果，本文增加了多车辆实验，如图 10 所示为输出的多车辆标定图像。通过图 8～10 的实验可以看到，本方法可以精确区分出车型，并清晰的标定出车辆轮廓，准确性与人工标定相近，但标定速度远超过人工标记。因此本方法在充分考虑车辆遮挡、环境复杂、目标小、种类多等因素后，本方法依然有较高的准确性及抗环境干扰能力。

图 8　输入视频截图

最后输入一段 2 小时 4 分钟的车辆视频，本文方法仅需要 3 小时 23 分钟即可生成 14 880 张车辆标定图像，人工标定需要 80 小时左右时间，标定速度相较人工标定提升

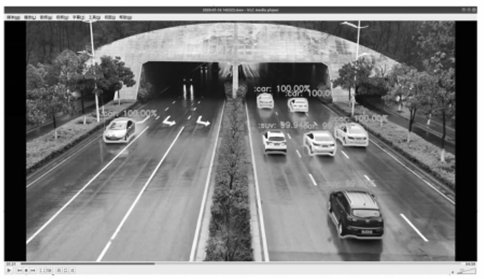


图 9 车辆数据标定图像

图 10 多车辆标定图像

95.77%。本文方法在保证精度的同时，大幅度减少了标定时间。

4 结束语

针对目前现有人工标定方法时间冗长、效率低下且容易出错的问题，本文提出一种基于 Mask Scoring R-CNN 的高质量数据集快速自动标定方法。通过与 ResNet50 网络相结合、调整了不同超参数并与迁移学习结合，在保证目标识别精度的同时降低了一半的训练时间；然后，建立了一种数据集评分机制，在遮挡、环境复杂、目标小、种类多环境下依然提高了目标掩膜标定精度；最后，提出了高质量多目标数据标定方法，保证数据集质量的同时大幅度降低了标定时间。从实验结果可以看出，本文方法具有精度高、训练数据量少、环境适应性强和标定时间短的优点。

参考文献：

[1] GALA R，VERMA S，KUMAR U，et al. A survey of intelligent traffic light control systems [J]. International Journal of Computer Applications，2018，180 (21)：31－36.

[2] ODAT E，SHAMMA J S，CLAUDEL C，et al. Vehicle classification and speed estimation using combined passive infrared/ultrasonic sensors [J]. IEEE Transactions on Intelligent Transportation Systems，2017：1－14.

[3] HILAL T，KIM G S，KIL T C. Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network [J]. IEEE Access，2018，6：2220－2230.

[4] HIDAYATULLAH P，FEIRIZAL F，PERMANA H，et al. License plate detection and recognition for indonesian cars [J]. International Journal on Electrical Engineering & Informatics，2016，8 (2)：331－346.

[5] SHIN J，SUNWOO M. Vehicle speed prediction using a markov chain with speed constraints [J]. IEEE Transactions on Intelligent Transportation Systems，2018，20 (9)：3201－3211.

[6] 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.

[7] 彭 博，蔡晓禹，唐 聚，等. 基于改进 Faster R-CNN 的无人机视频车辆自动检测 [J]. 东南大学学报（自然科学版），2019，49 (6)：1199－1204.

[8] 陈 辰，黄 晁，孙 松，等. 多模型融合车辆检测算法 [J]. 计算机辅助设计与图形学学报，2018，30 (11)：2134－2140.

[9] 袁功霖，侯 静，尹奎英. 基于迁移学习与图像增强的夜间航拍车辆识别方法 [J]. 计算机辅助设计与图形学学报，2019，31 (3)：467－473.

[10] SHAOQING R，KAIMING H，ROSS G，et al. Faster R-CNN：towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Itelligence，2017，39 (6)：1137－1149.

[11] PENGJIE T，HANLI W，SAM K，et al. G-MS2F：GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition [J]. Neurocomputing，2017，225 (15)：188－197.

[12] YEBES J，BERGASA L，GARCIA G，et al. Visual object recognition with 3D-aware features in KITTI urban scenes [J]. Cenes，2015，15 (4)：9228－9250.

[13] ZHENMING H，WEIBO F，TONG G，et al. A novel method based on a Mask R-CNN model for processing DPCR images [J]. Analytical Method，2019，11 (27)：3410－3418.

[14] KAIMING H，GEORGIA G，PIOTR D，et al. Mask R-CNN [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2018，37 (12)：2663－2673.

[15] KIM W，JUNG W，CHOI H，et al. Lightweight driver monitoring system based on multi-task mobilenets [J]. Sensors，2019，19 (14)：3380－3398.

[16] PHILIPPE X，FRANCK D，JEAN B，et al. Multimodal information fusion for urban scence understanding [J]. Machine Vision & Application，2016，27 (3)：331－349.

[17] 温尧乐，李林燕，尚欣茹，等. 一种改进的 Mask RCNN 特征融合实例分割方法 [J]. 计算机应用与软件，2019，36 (10)：130－133.

[18] 陈 敏，王 君，董明利，等. 改进的 Mask R-CNN 多尺度实例分割算法研究 [J]. 激光杂志，2020，41 (5)：40－44.

[19] 韩 进，刘恩爽，荣文忠. 基于全卷积网络的多车辆实时跟踪模型 [J]. 中国科技论文，2021，16 (11)：1234－1240.

[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.