单位代码 10635
学　　号 1120203340022246

西南大學

硕士学位论文

基于改进 EfficientDet 的车辆部件检测与重识别研究

论文作者：谢永生
指导教师：叶明
学科专业：计算机科学与技术
研究方向：计算机视觉
提交论文日期：2023 年 6 月 6 日
论文答辩日期：2023 年 5 月 27 日
学位授予单位：西南大学

中 国 • 重 庆
2023 年 6 月

Southwest University
Master's Thesis

Research on Vehicle Component Detection and Re-identification Based on improved EfficientDet Algorithm

Author Name：Xie Yongsheng

Supervisor：Ye Ming

June/2023

目　录

摘要

随着计算机视觉技术与深度学习的不断发展，基于深度学习的车辆部件检测与重识别成为研究热点。车辆部件检测和重识别是计算机视觉技术的一个重要分支，其在交通管理、智能停车、安防监控和自动驾驶等诸多研究领域有重要作用，同时在应用方面有宽广前景和切实价值。然而车辆部件尺度变换、视角和光照等因素给车辆部件检测与重识别带来了巨大的挑战。因此，有效提取图像中鲁棒的更具有辨识度的细节特征是提升车辆部件检测与重识别效果的关键所在。

本文提出了一种高效准确的车辆部件检测网络 EFDet-SPP，这个方法是基于深度学习网络 EfficientDet 构建的。首先，受到跨层连接以及残差网络的思想启发，设计了一种纵向交叉跨层连接的 BiFPN 网络进行特征融合，综合平衡了高层结点和低层结点的输入数据流，缩短了底层特征图到顶层特征图的距离，实现了更好的信息交互。同时将基于锚框预测转变为基于像素预测来克服模型应用场景变化带来的干扰，这样消除了与锚框相关的超参数从而减少了计算量，一定程度上提高了在车辆部件检测场景下的适用性。为了解决车辆部件均为小物体这一难题，在数据输入时，结合了 Mosaic 和复制粘贴两种数据增强方式来平衡样本，增强网络泛化能力。并在特征提取网络引入了空间金字塔池化来串联特征有效的捕获高语义信息。由于公共数据集的缺少，本文建立了两个车辆部件检测数据集 VLC 和 VDC，确保数据充足，场景丰富。

此外，本文聚焦于寻找具有区分度的局部特征，并不忽略整张车辆图像的全局信息，提出了一种基于部件与全局特征的多粒度车辆重识别算法。使用简化的 EFDet-SPP 捕获部件图像来提取局部特征，同时对整幅图像进行多尺度特征提取全局信息。这样不仅学习到了车辆的全局特征，也对区分度大的部件局部特征进行了有效的学习。更重要的是，本文针对更加细粒度的特征，设计了一种新颖的特征学习策略，即将判别信息划分为不同大小的粒度特征，同时不在语义区域进行学习，而是从垂直方向将图像特征均匀地划分为若干条纹，通过调整各个局部分支里的条纹数目，从而实现多粒度的局部特征表示。

为了测试本文提出的方法，在 VLC 和 VDC 数据集上进行了对比实验，证明了 EFDet-SPP 可以实现高效准确的车辆部件检测；在公共数据集 VeRi-776 和 VehicleID 上进行了车辆重识别实验，最后的实验结果验证了本文所提出的方法在车辆部件检测和重识别任务中的有效性。

关键词：车辆部件检测，车辆重识别，特征融合，局部特征，深度学习

ABSTRACT

Vehicle component detection and re-identification using deep learning has become an interesting research field as computer vision technology and deep learning advance. This area plays a crucial role in theoretical research in fields such as traffic management, intelligent parking, security monitoring, and autonomous driving. It holds broad prospects and practical value in application fields. However, vehicle component detection and re-identification are challenging tasks due to the scale transformation of vehicle components, viewing angle, and illumination. Hence, effective extraction of robust and recognizable detail features in images is key to improving the detection and re-identification of vehicle parts.

This thesis proposes an efficient and accurate vehicle component detection network, EFDet-SPP, based on the deep learning network EfficientDet. Firstly, a BiFPN network with vertical cross-layer connection is designed for feature fusion, inspired by the idea of cross-layer connection and residual network. This comprehensively balances the input streams of high-level nodes and low-level nodes, shortens the distance from the low-level feature map to the top-level feature map, and realizes better information interaction. At the same time, anchor box prediction is transformed into pixel prediction to overcome interference caused by changes in the model application scene, eliminating hyperparameters related to the anchor box and reducing the calculation amount, which improves applicability in the vehicle component detection scene to a certain extent. To tackle the challenge of tiny vehicle components, the combination of mosaic and copy-paste data enhancement methods balances the samples and improves the network's generalization ability. High semantic information is captured effectively by using spatial pyramid pooling in the feature extraction network. Due to the lack of public datasets, we have established two vehicle component detection datasets VLC and VDC to ensure sufficient data and rich scenes.

Moreover, this thesis focuses on finding local features with discrimination without ignoring the global information of the whole vehicle image. A multi-granularity vehicle re-identification algorithm based on component and global features is proposed. Local features are extracted from component images captured by the simplified EFDet-SPP, and global information is extracted from the whole image by performing multi-scale feature extraction. By doing this, global features of the vehicle and local features of the components with high discrimination are both effectively learned. More importantly, we

design a novel feature learning strategy for more fine-grained features, where the discriminant information is divided into granular features of different sizes. At the same time, instead of learning in the semantic region, the image features are evenly divided into several stripes from the vertical direction, and the number of stripes in different local branches is changed to obtain a multi-granularity local feature representation.

To test the proposed method, we conducted comparative experiments on datasets VLC and VDC, proving that EFDet-SPP can achieve efficient and accurate vehicle part detection. We used the public datasets VeRi-776 and VehicleID to conduct vehicle re-identification experiments. The final experimental results verified the effectiveness of the proposed method for vehicle component detection and re-identification tasks.

Keywords: Vehicle component detection, Vehicle re-identification, Feature fusion, Local feature, Deep learning

第 1 章 绪论

1.1 研究背景及意义

公安部公示的数据显示，到2021年11月底时，全国机动车保有量已达到3.93亿辆，其数量是十年前的 1.64 倍[1]，车辆交通已经成为现代社会的重要标志之一。车辆发展既有利又有弊，它给人们的生活与工作带来便利的同时也引发一系列的社会与环境问题。比如城市交通堵塞、交通事故频发和套牌违法、汽车被盗、交通肇事逃逸以及利用汽车作案等案件呈年年增加趋势。这些问题已经成为当前世界各国所面临的共同问题，为了克服这些问题，智能交通系统[2]将多种先进的科学技术整合在一起，有效地应用于车辆和交通的监管中，这样显著减少了交通系统中存在的各种问题。

智能交通系统的核心功能是检测车辆位置和各部件的定位。最近，视频图像处理技术的发展推动了基于图像的车辆检测、重识别技术在智能交通系统中的应用[3]。一方面，我国越来越重视道路监控，视频设备的成本降低，使得通过视频监控收集车辆图像数据成为主要的数据采集方式。另一方面，在诸如汽车生产企业、车管所和检测站等需要安全检查的地方，车辆安全部件检测是极其重要的一环。此任务不仅要识别出车辆，而且需要更为精确的识别轮胎、安全带等零部件还有灭火器、三角警示牌等安全装置。对于安全检查员而言，采用高精度的实时检测器可以显著提升其在车辆外观检查、尾气检测和照明检测等方面的工作效率。在车辆搜索系统中，车辆部件检测是至关重要的环节。车牌号码虽然是车辆的唯一标识，但是由于套牌、遮挡和缺失等违法行为的出现，仅依靠车牌难以准确匹配目标。除车牌外，车标和轮胎等部件也含重要信息，稳定且难以更改，可为查找车牌资料丢失的车辆提供重要途径。结合上述三点分析，基于图像的车辆部件检测研究形成了一项价值巨大的研发课题，其成果有助于缓解我国目前拥堵的交通环境、预防车辆潜在的安全隐患、提高公安部门的犯罪侦查效率等，对智能交通系统未来的发展具有重要的战略意义。此外，可以为我国城市化进程和交通出行需求的满足做出贡献，具有重大的社会和经济价值。

“智能交通”、“智慧城市”在人们的生活中越来越普遍，车辆重识别技术是智能交通系统中的一个重要技术应用，这项技术旨在一系列复杂的非重叠场景中，通过筛选和识别，找到不同摄像头拍摄的同一辆车辆图像。任务示意图如图 1-1。车辆重识别技术主要用于追踪、车辆违章、智能交通管理和无人驾驶。首先，在对车牌信息错误或缺失的犯罪嫌疑车辆进行追踪时，使用多摄像头追踪特定目标，若相机的位置已知，可用车辆重识别系统描绘其移动路径。其次，交警在进行车辆违章判罚时，可利用摄像头捕获到的车辆图像来判断目标车辆是否违反交通规

则，例如闯红灯、压线、错误行驶车道等。然后，在智能交通管理方面，越来越多商场、旅游景区的停车场配备了自动收费系统，需对无牌车、套牌车和车牌被遮挡的车辆进行车辆重识别来匹配以计费。最后，行车记录仪可视为无人驾驶车辆上的移动摄像头，可捕获到不同角度的图像。因此，利用车辆重识别技术可以帮助分析一段时间内无人驾驶车辆周围车辆的行动轨迹，有助于推进无人驾驶技术的研究与完善。

图 1-1 车辆重识别任务示意图

Fig.1-1 Example of vehicle re-identification task

综上所述，车辆部件检测与重识别在车辆安全检测、检索与匹配等领域起到了极其重要的作用，同时也可以解决车辆交通流量管理、车辆环境监测和信息获取等方面的问题。因此车辆部件检测与重识别在理论研究方面发挥着重要的作用，同时在诸多应用领域有宽广前景和切实价值，相信未来将会应用到越来越多的现实场景下。

1.2 国内外研究现状

20 世纪以来，专家们在车辆检测与重识别领域进行了大量的研究，发表了许多显著的成果。伴随着深度学习的兴起，给车辆部件检测与重识别领域带来了巨大的变化，越来越多的基于卷积神经网络的算法被提出。车辆部件检测属于目标检测的子问题，因此车辆部件检测同样可以使用目标检测的相关方法。本节将介绍目标检测、车辆部件检测和车辆重识别的国内外研究现状以及现在所面临的问题与挑战。

1.2.1 目标检测研究现状

目标检测是计算机视觉中一项重要的任务，同时也是具有挑战性的，需要对图像中的目标进行分类和定位。二十年以来，目标检测的发展经历了深度学习出现之前和深度学习出现之后两个阶段。在深度学习出现之后，高效准确的基于深度学习的目标检测算法迅速超越了传统算法，成为主流。基于深度学习的目标检测算法可被分为有锚框（Anchor-based）检测模型和无锚框（Anchor-free）检测模型。同时前者由两阶段目标检测和单阶段目标检测两种方式组成。

对于 Anchor-based 检测模型，可分为基于候选框的两阶段检测算法和基于回归的单阶段检测算法，这两类目标检测算法的一般框架如图 1-2 所示。对于两阶段目标检测器，首先通过一些图像分割算法生成一组稀疏的候选区域，然后采用滑动窗口的思想，对卷积神经网络对每个区域进行分类。它最初在 Girshick[4]提出的 RCNN 中被介绍推广，是首个将卷积神经网络引入目标检测的算法；Girshick [5]等人提出了 Fast R-CNN 算法，其中卷积层的卷积是直接对整张图像，而不是对每个候选区域进行，这样减少了大量重复计算提高了速度和准确性；Ren[6]等人提出的 faster RCNN 引入了区域提议网络（Region Proposal Networks，RPN），这是一个全卷积神经网络，可以同时预测每个物体位置的区域边界和对象分数。后来，为了提高目标检测算法的性能，研究者们提出了很多算法，包括架构重新设计[7,8]、注意力机制[9]和特征融合[10]等。

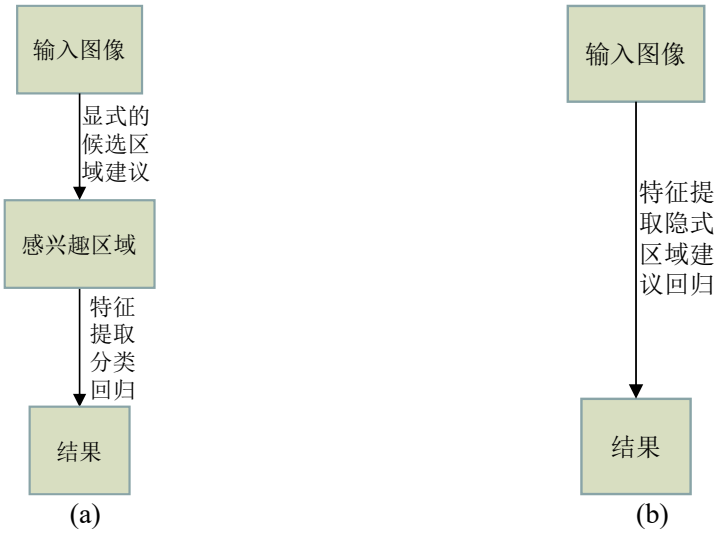


图 1-2 目标检测算法框架。 (a) 两阶段目标检测算法框架；(b) 单阶段目标检测算法框架

Fig.1-2 Object detection algorithm framework. (a) Two-stage object detection algorithm framework; (b) One-stage object detection algorithm framework

另一种方式就是单阶段目标检测，其将目标的分类和定位问题直接转换为回归问题。现如今的单阶段目标检测器性能是惊人的，它们能快速、高效地预测物体从而引起了广泛的关注，属于这个类别的网络有 YOLOv2、YOLOv3 和 YOLOv4[11-13]。YOLOv2 包括批量规范化、高分辨率的分类器、通过锚框预测的边界框，以及一个名为 Darknet19 的自定义特征提取网络。在 YOLOv3 中，网络预测是通过三个不同的尺度进行的，这使得网络比其早期版本更加有效，同时采用 Darknet53 来进行特征提取，将逻辑回归代替 softmax 层。YOLOv4 论文解释了改善目标检测网络的各种方法的准确性和效率。它使用了多种新的功能，如加权连接（Weighted-Residual-Connections，WRC）、交叉小批归一化（Cross miniBatch Normalization，CmBN）、DropBlock 正则化和 CIoU 损失来增强网络性能。还有 Liu W[14]等人提出的 SSD，基于检测转化为回归的思路，可以一次完成目标定位与分类；谷歌的 Brain Team[15]提出了 EfficientDet，一个可扩展的高效单阶段检测器，其提出了一种加权双向特征金字塔网络（BiFPN），可以实现简单快速的多尺度特征融合。

对于 Anchor-free 检测模型，这些方法放弃了锚框并使用像素作为训练样本直接分类和锚定对象。最早的 YOLOv1[16]将目标检测作为一个空间分离的边界框和相关的类概率的回归问题。然后 CornerNet[17]利用检测锚框的一对角点来识别出目标，同时提出角点池化（Corner Pooling），来更好的定位锚框的角点。这个框架需要更复杂的后处理来对同一对象的角进行分组。CornerNet-Lite[18]是 CornerNet 两种有效变体的组合，均可以提高速度。为避免复杂的处理步骤，CenterNet [19]直接预测对象的中心键点。ExtremeNet [20] 是一种基于关键点的目标检测算法，它使用标准关键点估计框架，通过对每个目标类预测 4 个多峰值的热力图来寻找极值点，同时使用每个类的热力图来预测目标中心，它的性能比 YOLOv3 更好，而且速度更快。RepPoints[21]将目标表示为一组样本点，并将它们限制在目标的空间范围内。与前述方法相比，FCOS [22]具有相对简单的结构，将图像内的所有像素视为训练样本，直接预测像素到一个边界框的四个距离。同时，为了抑制低质量预测边界，引入了一种新颖的“Center-ness”得分和回归分支并行。

1.2.2 车辆部件检测研究现状

如果仅仅确定车辆的位置和类别，可以分为三类：基于运动的车辆检测算法、基于知识的车辆检测算法和基于深度学习的车辆检测算法。在基于运动的车辆检测算法中，背景建模通常要求相机静止，然后使用混合高斯模型去除背景信息，之后通常可以进行一些图像处理操作获取运动目标，相对简单，在目标运动明显

时甚至优于很多复杂的机器学习方法。另一种典型的基于运动的车辆检测方法就是光流法。高磊[23]采用 Harris 算子计算图像的特征点，然后计算特征点的光流，最后使用聚类的方式提取车辆目标。对于基于知识的车辆检测算法，利用车辆外观的明显特征来识别图像中的车辆，这些外观特征包括车辆的对称性、车辆底部的阴影等。Hilario C H[24] 利用边界算子计算图像的边缘信息，然后利用车辆尾部对称的局部特征检测车辆。邬紫阳[25]利用车辆在道路上行驶底部会存在阴影这一知识来定位车辆可能存在的区域，再使用其他后处理方法来检测出车辆。随着技术的发展，后来出现了一些比以前的边缘和局部对称特征更可靠的特征描述符。文献[26-28]采用了面向梯度直方图方法提取图像中的车辆类型特征，利用支持向量机（SVM）对特征进行分类，从而实现车辆检测。Pan C[29]等人提出了可变形零件模型（DPM），其是一种用于车辆检测的方法，取得了良好的效果。不过前两种方法具有极大的局限性，都是从整车出发进行检测，如果想对车辆的部件进行识别检测，基于深度学习的车辆检测算法更具有可靠性。

不管哪类方法，具体到车辆部件检测的相关研究比较少。Alberto[30]利用 Viola 和 Jones 提出的基于 Haar-like 特征的级联分类器来车辆部件检测，包括车轮、后灯、后视镜、车窗、侧面车的保险杠等；Brian Leung[31]采用基于可度量特征的 Gentle Adaboost 检测车轮、车灯、天窗、后保险杠和后视镜等车辆部件。这些方法均采用滑动窗口模型和手动设计特征，因此其计算复杂，识别效果不准确，基于深度学习的目标检测有极大的优势，故本文主要研究基于卷积神经网络的车辆部件检测算法。

1.2.3 车辆重识别研究现状

车辆重识别属于目标检测的一种，随着行人重识别的流行，车辆重识别的研究逐渐引起关注。车辆重识别任务可以看作实例图像检索，任务难度较大。现有的车辆重识别算法主要分为两大类：传统的车辆重识别算法和基于深度学习的车辆重识别算法。

传统的车辆重识别算法主要集中于使用特征工程来人工细化和清理数据。一般来说，特征提取方法有三种：尺度不变特征变换[32]（Scale-Invariant Feature Transform，SIFT）、定向梯度[33]（Histogram of Oriented Gradient，HOG）和局部二值模式[34]（Local Binary Pattern，LBP）。除这三种方式之外，众多学者还做了很多有意义的研究，如自旋图像[35]、加速鲁棒特征[36]（Speeded Up Robust Features，SURF）、时空兴趣点[37]（Space-Time Interest Points，STIP）和运动边界直方图[38]（Motion Boundary Histogram，MBH）。然而，这种方法只对特定任务有效，不能适应不同的应用场景，例如颜色直方图特征：它对图像分类任务有效，

但对语义图像分割没有效果。基于手工制作的特征只关注图像的某些特点，如HOG关注图像的边缘信息，LBP关注图像的纹理等。换句话说，这些特征只能在训练数据上表现良好，而在新数据上的表现可能会很差，它们的泛化能力较差。

卷积神经网络（Convolutional Neural Network，CNN）被提出以后，弥补了特征工程提取手工特征的不足，同时也更适合图像检索任务。在学习过程中，此类方法通常旨在构建识别汽车的神经网络模型。经过训练，网络可以将车辆的图像映射到特征空间，并使用特征空间中的相似性度量 （例如欧几里德距离）来表示车辆之间的距离。如果车辆图像之间的特征距离较小，将被视为同一车辆。此方法一般侧重于开发有效的损失函数来提高网络训练效率。例如，大间隔余弦损失[39]（Large Margin Cosine Loss，LMCL）的目标是最大化类间方差和最小化类内方差；三元组损失[40]（Triplet Loss）的目标是通过优化三个硬样本之间的距离来学习视觉表示；Y. Zhang[41]等人提出一种引导型三元网络结构，其在原来三元组损失函数的基础上加入了分类损失，以用来限制原来的训练网络，从而提高了重识别的效率。Sun 等人[42]提出了圆损失（Circle Loss），它自适应地调整每个相似性分数的权重。在重识别训练中，采样策略也起着至关重要的作用。分层三重损失[43]（Hierarchical Triplet Loss，HTL）定义一个编码上下文信息的层次树来收集信息样本，这有助于克服训练三重损失时随机抽样的局限性。然而，采样策略通常是启发式的，取决于损失函数，并且很难调整。除此之外，后处理对于减少假阳性预测也很重要。例如，重排序可以提高排名列表的准确性，其通常依赖于初始重排序的图库图像的一致性和最近邻关系。Zhong[44]等人提出 k-互反编码方法，该方法考虑了两幅图像之间的原始距离和焦距距离。本文将利用车辆部件检测算法来提取车辆的局部特征并结合其全局特征来进行车辆重识别的研究。

1.2.4 现阶段的研究挑战

深度学习发展迅速且在每个领域都大放异彩，但是在不同的应用场景中，车辆部件检测和重识别的准确率常不尽如人意。首先对于车辆部件检测问题，存在如下难点与挑战：

（1）数据集与域适应问题

深度学习是以数据为驱动的，场景充足、高质量的数据集是车辆部件检测发展的关键一环。在车辆部件检测中，公共数据集往往是对整车的检测，当下还没有一个公开完整的车辆部件图像数据集，这样造成的结果是深度学习算法的准确性下降、泛化能力不足。汽车部件的种类繁多、收集困难和采集车辆部件数据涉及的隐私问题均导致了目前可用的车辆部件数据集非常有限，这使得训练准确的识别模型变得更加困难。

跨领域目标检测存在大量困难，常因为两个场景的不同严重降低网络的性能。公共数据集往往是固定视角，拍摄场景比较固定，光照视角的变化会影响网络的泛化能力与鲁棒性。同时表现较好的目标检测器，应用在专业车辆部件领域的效率和精确度差强人意。

（2）小目标检测问题

车辆部件大部分都是小目标，其信息蕴含在固定尺寸的图像里。对于图像的不断卷积提取特征，使得最终得到的特征图分辨率较低，并且小目标本身的像素就较少，这样会导致小目标高语义信息不足，同时这些小目标的蕴含信息也可能在不断降采样中逐渐模糊。小目标形状各异，往往与背景混合在一起，有时也会与其他物体发生拥挤和遮挡问题。这些问题往往使得检测网络的性能低下、准确率不高，给车辆部件检测带来了一定的局限性和挑战性。

随着行人重识别被关注，车辆重识别也发展迅速，在许多公有测试集上效果很好，但是在现实应用中仍旧存在很多挑战：

(a) (b) (c)

图 1-3 车辆重识别难点问题。 (a) 车辆外观相似；(b) 分辨率低下；(c) 尺度变化

Fig.1-3 Difficulties in vehicle re-identification. (a) Similar vehicle appearance; (b) Low resolution; (c) Scale variation

（1）不同车辆外观相似

对于同一厂家不同车型的汽车，他们的外观结构设计极为相似，有时关键性的区别信息（如车牌、车标）可能会被遮挡而无法产生效果带来了一定的局限性。更有甚至对于同一品牌的同一型号的车辆在出厂时几乎是完全一样的，这是车辆重识别任务中的最大挑战。如图 1-3(a) 所示，两辆黑色轿车外形完全一样，单纯依靠肉眼都难以辨别，这样的车辆往往被断定为同一辆车。

（2）分辨率过低

在实际场景中，监控性能不足导致捕获的图像分辨率低下。这样在使用神经网络提取特征时，使车辆特征越来越模糊，往往提取不到有效的特征。如图 1-3(b) 所示，仅仅是图像都是模糊难以区别。

（3）尺度变化

摄像头拍摄高度与距离是与拍摄图像的大小相关联的，近距离拍摄的车辆较大，反之亦然。同时在监控系统中，摄像头拍摄高度和距离往往是存在差异的，这样就会存在尺度不一的问题。这导致重识别网络提取的特征信息量不同，最终难以匹配。如图 1-3(c) 所示，尺度变化带来图像信息量的不同。

1.3 本文主要工作

本文对车辆部件检测与重识别领域存在的难点和问题，提出了不同方向的优化与改进，并且通过对比试验证明其有效性。具体地，从车辆部件检测和车辆重识别两个方面来进行研究，故本文的主要工作为：

（1）在车辆部件检测中，针对当前公用车辆数据集有限，建立了车辆部件检测数据集，确保数据充足，场景复杂且具有普遍性。然后本文基于 EfficientDet 网络框架设计了一种高效准确的车辆部件检测网络 EFDet-SPP。在数据输入，针对车辆精细部件这类小目标对像，结合了 Mosaic 和复制粘贴的数据增强方法对小样本数据进行扩充使样本达到平衡。针对图像输入尺寸偏小且小目标在特征提取的过程中像素信息逐渐不足问题，首先在特征提取网络引入了空间金字塔池化模块，对特征图进行多尺度采样，串联特征且有效地捕获高语义信息，然后在特征融合网络 BiFPN 上增加纵向的交叉跨层连接数据流，平衡了自下而上路径的特征结点的输入数据流，充分利用不同层级之间丰富的特征信息，这样在一定程度上解决了小目标检测较差的情况。同时为了增加 EfficientDet 在车辆部件检测领域的适应性，在分类定位网络中借鉴 FCOS 无锚框的检测方式，将基于锚框预测转变为基于像素点预测，消除了与锚框相关的超参数从而减少了计算量，一定程度上提高了在车辆部件检测场景下的适用性。最后通过对比实验完成了对本文算法有效性的验证。

（2）在车辆重识别中，借助着 EFDet-SPP 对车辆部件的优势，提出了一个基于部件与全局特征的多粒度车辆重识别算法。这是一个包含局部支路（车窗、车脸）和全局支路的三分支网络，在局部支路中，本文提出了多粒度学习的思想，设计了针对局部部件的细节特征处理模块。从垂直方向均匀地将图像特征分割成几条条纹，并改变不同局部分支中的条纹数，以获得多粒度的局部特征表示；在全局支路中利用尺寸不同的卷积核来提取多尺度特征，之后将其叠加获得全局特征。因此，这样不仅学习到了车辆的整体特征，也对区分度大的部件局部特征进行了有效的学习。同时对 EFDet-SPP 进行简化，构建了一个轻量的部件检测模块来提取部件图像。最后，在 VeRi-776 和 VehicleID 这两个公共车辆数据集上进行了实验证明了该算法具有较高的精确性。

1.4 论文组织结构

本文的研究主要分为两个部分：首先从车辆部件检测算法的精度和速度两个方向入手构建一个高效准确的无锚框车辆部件检测模型；其次多粒度提取车辆部件所包含的局部特征完成有竞争力的车辆重识别算法。基于上述研究工作，将各章节的组织结构安排如下：

第一章：绪论。绪论介绍了车辆部件检测与重识别的研究背景意义以及面临的挑战，同时综述了其国内外研究现状，最后详细阐述了本文对该领域做出的主要工作。

第二章：相关理论基础。本章主要介绍了车辆部件检测与重识别的所用到的相关理论基础，主要包括卷积神经网络的组成、车辆部件检测算法概述和车辆重识别概述等，并介绍了研究对象和任务流程、可以进行改进优化的技术以及进行评测常用的指标等。

第三章：优化 EfficientDet 的车辆部件检测算法。首先介绍了一种可扩展的高效单阶段检测器 EfficientDet，但是发现其对于车辆部件领域数据发挥并不好，尤其是小目标的检测。然后，针对性的进行了数据增强以及添加了空间金字塔池化模块和特征融合网络的纵向交叉跨层数据流。同时改变了其预测方式，从基于锚框改进为无锚框网络，转变为基于像素预测增加其在不同场景下的适用性。最后对提出的 EFDet-SPP 网络进行了测试。

第四章：基于部件与全局特征的多粒度车辆重识别算法。主要介绍了本文提出的多粒度车辆重识别网络设计，包括各个模块和分支的结构，首先在部件检测模块，阐述了对于 EFDet-SPP 的轻量模型，然后在此基础上，设计了多粒度学习策略，并进行多尺度特征提取，结合 softmax 损失函数和批量硬三元组损失函数优化来进行车辆重识别。最后对各个分支进行了对比实验，并与主流车辆重识别算法进行了比较验证了整个网络模型的有效性。

第五章：总结与展望。总结了全文对于车辆部件检测与重识别算法研究的主要内容；然后对未来的研究方向和计划做出展望。

第 2 章 相关理论基础

2.1 引言

车辆部件检测和重识别均与深度学习有着密不可分的关联，两者都依赖于卷积神经网络的特征提取能力。本章除了对卷积神经网络相关理论做简单的介绍之外，也将对后续将用到的目标检测，目标重识别技术进行分类介绍。本文用到的技术大部分是基于深度学习的算法，故主要从原理出发介绍了特征提取，车辆部件检测以及车辆重识别的工作内容和相关理论。

2.2 卷积神经网络

基本的卷积神经网络[45]可以认为是一种带正则化的多层感知机结构，其擅长解决图像尤其是大尺寸图像的机器学习问题，因此卷积神经网络在视觉领域有广泛的应用。卷积神经网络主要被用于提取图像特征，进一步分析视觉图像，与传统特征提取方式相比，其不需要人工设计特征。在 2010 年之前，由于缺乏数据集、计算资源，CNN 发展较为缓慢，不过随着深度学习的发展以及 GoogleNet[46]、VGG[47]、ResNet[48]和 DenseNet[49]等具有强大图像特征学习能力模型的提出，CNN 因其强大的图像泛化能力，引起了国内外众多研究者的关注。

2.2.1 卷积神经网络的基本组成

卷积神经网络通过一系列的卷积和池化操作对输入数据进行下采样，减少数据的维度以及计算量，最终使其能够被训练。图 2-1 是一种简单的卷积神经网络的主要结构，通常由输入层、隐含层和输出层三个部分组成。其中输入层将数据送入 CNN 进行特征提取；输出层负责将隐含层的输出映射到目标空间，得到最终的输出结果；隐含层负责核心工作：经过隐含层运算以后，输入图片将被映射成高维特征获得语义信息，最后得到更抽象化的特征。隐含层主要由卷积层、池化层、激活层和全连接层 4 个基本单元所构成，这些基本单元分别有着不同的作用。

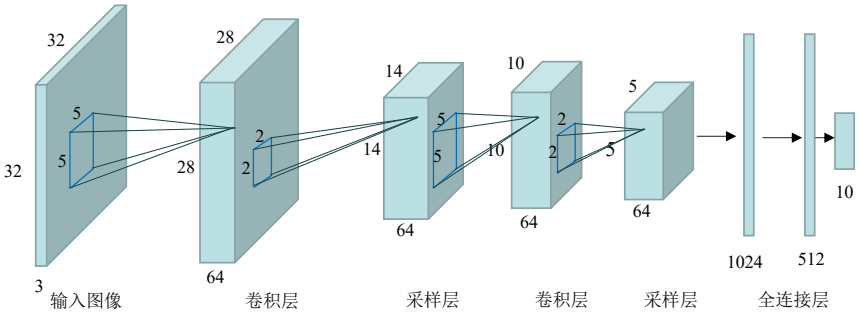


图 2-1 简单卷积神经网络

Fig.2-1 Simple convolutional neural network

（1）卷积层

卷积层（Convolution Layer）是卷积神经网络中非常重要的一层，通过滤波器对输入数据进行卷积运算并生成特征图，实现特征提取。卷积运算的目的是提取输入图像的各种特征，如边缘、线条和角点。因此图像和卷积核可视为矩阵，图像和卷积核在每一个位置通过相乘和相加的操作来生成特征图。所以可以把卷积表示为：

$$y_j^l=f\left(\sum_{i\in M} y_j^{l-1}\times k_{i,j}^l+b_j^l\right) \tag{2-1}$$

式中，y_j^l 为第 l 层的输出特征层；f 是激活函数；M 是特征图的位置集合；$y^{l-1}{}_j$ 为前一层的输出特征作为第 l 层的输入；$k^l_{i,j}$ 是 j 位置卷积核对应 i 位置特征图的权重；b^l_j 是逐通道的偏置项。

卷积层的参数主要有卷积核大小（Kernel Size，k），步长（Stride，s），填充（Padding，p）以及输入输出的宽度（Width，W）、高度（Height，H）和通道数（Depth，D）。

卷积核：卷积层覆盖的范围被称为感受野，感受野与卷积核尺度密切相关，卷积核尺度越大，感受野越大，信息提取效率越高。卷积核的一般尺寸大小有 3×3（如图 2-2）、5×5 和 7×7，当然在不同的场景任务下有不同的需求，要根据网络要求去使用不同尺寸的卷积核，甚至于 1×1 大小的卷积核。

步长：卷积核在图片上移动遍历每一个像素，每次移动的大小就是步长。卷积核的步长表示图像特征提取的精度，通常情况下，步长被设置为 1，在图 2-2 中，卷积操作的步长设置为 1。除此以外，也可以使用步长为 2 的下采样卷积，其输出特征图尺寸以及所包含的图像信息将会变小。

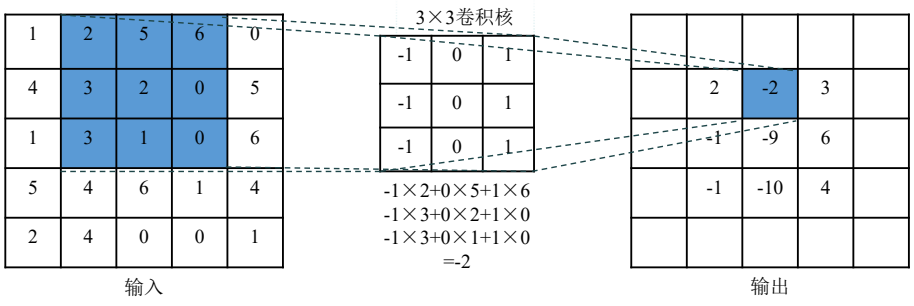


图 2-2 卷积过程

Fig.2-2 Convolution process

填充：如图 2-2 所示，图像输入尺寸：5×5，卷积核大小：3×3，卷积核大小和输入图像并不匹配，如果直接卷积（假定步长为 1），最后得到一个 3×3 的输出。这样做会导致每次做卷积操作图像就会缩小，同时图像边缘信息无法完全

利用，意味着缺乏图像边缘的位置信息。为了解决上述问题，通常需要填充缺失的图像边缘区域，目前广泛使用的填充方法有两种：Valid 卷积和 Same 卷积。Valid 卷积不会填充像素，这样的话，假如有一个 $n\times n$ 的图像，使用 $k\times k$ 的滤波器做卷积，最后你会得到一个 $(n-k+1)\times(n-k+1)$ 的输出；Same 卷积表示只要选择相应的填充尺寸，就能确保得到和输入相同尺寸的输出。

输入输出的宽度、高度和通道数：每个卷积核都会对输入图像进行卷积操作，生成一个输出特征图。因此，如果有 n 个卷积核，那么输出特征图的通道数就是 n，卷积核的数量决定了输出特征图的通道数。如图 2-2 所示，卷积核在一个 $n\times n$ 的输入图像上根据步长移动，在每一个它所经过的位置与所覆盖的输入进行矩阵运算，最终将输出像素按相对位置排列而得到输出特征图。一张图片经过多次卷积，充分提取其包含的有效特征信息。输出特征层的尺寸（$W\times H$）与卷积核尺寸，步长和填充有关，计算公式如下：

$$W_{output}=\frac{W_{input}-k+2p}{s}+1 \tag{2-2}$$

$$H_{output}=\frac{H_{input}-k+2p}{s}+1 \tag{2-3}$$

其中 W_{input}、H_{input}、W_{output} 和 H_{output} 为输入输出特征图的宽度和长度；k 是卷积核尺寸，p 是填充数，s 是步长。

（2）池化层

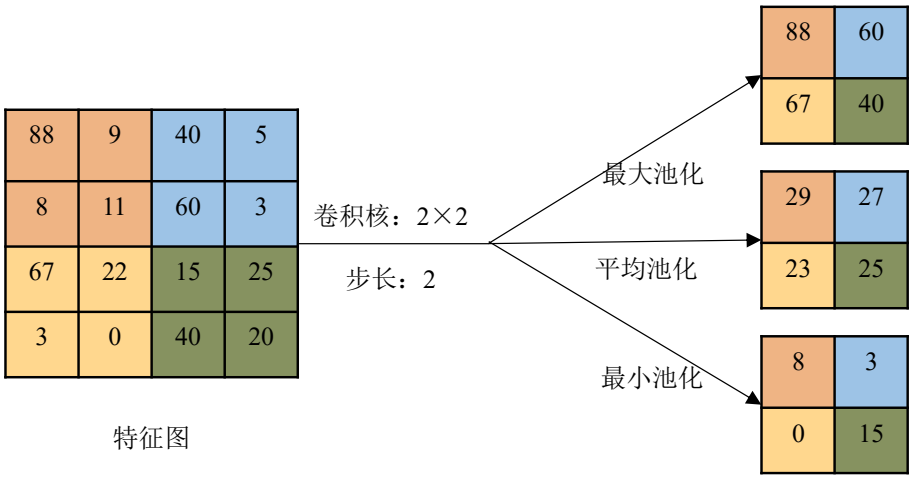


图 2-3 最大、平均和最小池化对比图

Fig.2-3 Comparison of max pooling， average pooling and min pooling

池化层（Pooling Layer）本质上是一种形式的降采样，主要用于降低数据维度和参数量，从而有效地减少过拟合的风险和计算量，常见的池化函数如图 2-3 所示，最大最小池化方式是取卷积核所覆盖特征图的部分矩阵进行取最大值或最

小值，而平均池化则是将这一部分进行平均，同时可以很容易发现池化操作缩小了特征图。池化层的引入是通过模拟人类视觉系统，对视觉输入对象进行降维、抽象。池化之所以有效的原因在于一旦定位了一个特征，其确切位置远不如其相对于其他特征的相对位置重要。总之，池化层的作用是保留主要特征的同时减少参数和计算量，防止过拟合。池化层的另一个作用是可以减少特征图的尺寸，从而减少最后连接层中的参数量。

（3）激活层

激活层（Activation Layer）用于解决卷积神经网络的非线性问题，如果只用卷积层，无非就是多个矩阵相乘，因此通常会在每个卷积层后面添加激活层，以引入非线性因素，从而使神经网络可以拟合非线性函数。如图 2-4 为激活层的结构图。在图中 x_1~x_n 是图像的输入特征；w_{1j}~w_{nj} 为输入特征与神经元 j 之间连接的权值；y_j 是经过激活函数的输出值；f 表示激活函数。通常来说，常用的激活函数有 Sigmoid、Tanh 和 ReLU[50]。

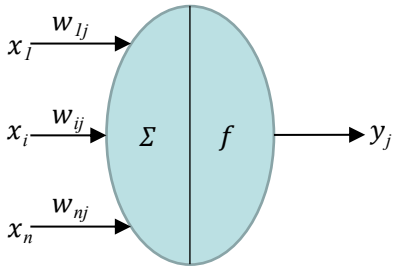


图 2-4 激活层结构图

Fig.2-4 Framework of the activation layer

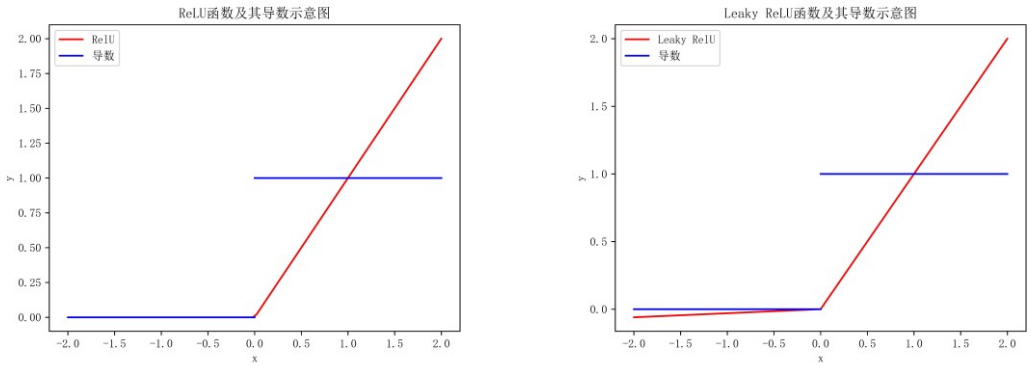
在激活函数中，Sigmoid 和 Tanh 函数图像较为相似均会在输入达到设定阈值后进入饱和区，同时两者都含有幂运算，计算比较耗时以及会造成梯度消失等问题。相比于前者，ReLU 函数在输入大于 0 时没有上界，因此不会导致梯度消失问题，使得可训练深层网络，其计算公式如下：

$$f(x)=\begin{cases} x, & x\geq 0 \\ 0, & x<0 \end{cases} \tag{2-4}$$

图 2-5(a) 是 ReLU 函数与其导数的图像，可以看到在负半轴函数图像与导数图像重合。其导数在正负半轴均保持稳定，所以在训练中也更加稳定，但是当输入值为负的时候，输出始终为 0，其一阶导数也始终为 0，这样会导致神经元不能更新参数，也就是神经元“死亡”问题。新的激活函数 Leaky ReLU[51]被提出来解决此问题，图 2-5(b) 是 Leaky ReLU 的函数以及导数图像，其表达式如下：

$$f(x)=\begin{cases} x, & x\geq 0 \\ \alpha x, & x<0 \end{cases} \tag{2-5}$$

式中，α 的值通常设置为 0.01 左右，Leaky ReLU 的优点是在输入为负数时，不会直接输出 0，而是输出一个斜率较小的线性函数来解决神经元“死亡”问题。



(a) (b)

图 2-5 激活函数示意图。 (a) ReLU 函数及其导数图像；(b) Leaky ReLU 函数及其导数图像

Fig.2-5 Activation function image. (a) ReLU function and its derivative image; (b) Leaky ReLU function and its derivative image

（4）全连接层

全连接层（Fully Connected Layer）每个神经元都与上一层的所有神经元相连，因此其参数量很大，需要大量的计算资源。全连接层通常用于分类任务，例如对于一个二分类问题，输出结果是一个 2×1 的向量，其中每个值对应各个类别的得分，卷积、池化和激活等层次可以将输入图像数据转换为隐层特征，而全连接层则负责将这些隐层特征映射到样本标记空间，最终用于分类器的输入。

2.2.2 经典网络模型

21 世纪以来，国内外学者为了进一步开发卷积神经网络强大的计算性能，提出了许多经典的网络结构模型来解决各种深度学习问题（如目标检测、图像分割和车辆重识别）。

AlexNet[52]是 Geoffrey Hinton 引入的首个深度架构，其在 2012 年的 ImageNet 比赛中大获全胜，将错误率降低了 10 个百分点，成为了深度学习的里程碑。它是一个简单而强大的网络架构，其包含 5 个卷积层和 3 个全连接层，激活函数使用 ReLU，同时采用 Dropout[53]随机丢弃神经元以防止过拟合。该模型的独特之处在于它的规模和使用 GPU 进行训练，通过使用 GPU 将训练速度提高了 10 倍。

自 AlexNet 问世以后，许多学者在小卷积核和多尺度两个方向来改进网络结构以提高模型的准确率，VGG[47]网络的作者则选择了加深网络深度这一方向。其网络架构基本上均使用的是 3×3 的卷积和 2×2 的最大池化，这样在增加网络深

度的同时减少了网络参数量。同时少部分采用了 1×1 卷积，这样可以增加网络的非线性，提高了网络的表达能力。图 2-6 是具有较好性能的 VGG-16 的网络架构，一共包含了 13 个卷积层和 3 个全连接层。该模型输入图像的大小为 224x224，经过若干卷积、池化和全连接层得到预测结果的概率分布。具体来说，它包括多个卷积层，其中每个卷积层使用不同数量的卷积核，随后进行最大池化。最后，通过三个全连接层并使用 softmax 函数来得到预测结果的概率分布。

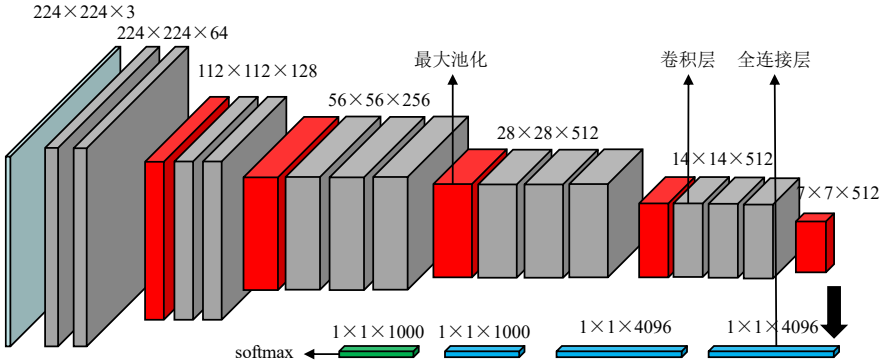


图 2-6 VGG-16 网络架构示意图

Fig.2-6 Illustration of VGG-16 network architecture

VGG 被提出之后，国内外研究人员致力于使网络变宽变深，因为理论上随着神经网络层数的增加，网络可以进行更复杂的特征提取，但实际训练中难以训练，梯度消失会导致准确度饱和甚至下降。残差神经网络（ResNet）[48]提出一个深度残差模块来来解决这个问题，通过短路机制引入残差单元并提高网络深度，同时使用步长为 2 的卷积进行下采样，最后用全局平均池化代替全连接层。这一改进使得梯度可以在深层网络中得到良好的传递，将网络深度提升到 152 层。在深度学习任务中，ResNet 在深度方面为网络结构提供了可行性解决方案，在网络优化方面也给出了更大的可改善空间。

2.3 车辆部件检测概述

2.3.1 研究对象定义

基于深度学习的车辆部件检测任务的目标是在整张图像中找到感兴趣目标的位置，同时根据特征信息正确识别出对应的类别，简单来说可被分解为定位和分类两大主要任务。其中车辆部件就是需要定位的目标，车辆部件包括车架号、铭牌、轮胎、左右车灯、后视镜、排气孔、车牌和车标等。这些车辆部件的类型即是在分类任务中需要判断的，车辆部件类型的信息是先验知识，由研究初期制定方案所决定。

2.3.2 锚框

车辆部件检测中，一般用边界框来表征目标即部件的相对空间位置，边界框通常是矩形的，其表示方式有两种：矩形左上角的坐标（x_1，y_1）以及右下角的坐标（x_2，y_2）表征和边界框的中心坐标以及框的宽度和高度（x，y，w，h）表征，两种不同的边界框表示方式是可以相互转换的。在目标检测算法中，通常以是否使用锚框为根据将其分为 Anchor-based 和 Anchor-free 两种目标检测算法。如图 2-7 所示，锚框就是以锚点为中心，经过缩放比和宽高比这些预定义的超参数计算生成多个不同的边界框，用于检测不同大小和长宽比的目标。对于图中的锚点来说，设置了三种不同的长宽比，因此共生成了 3 个锚框。

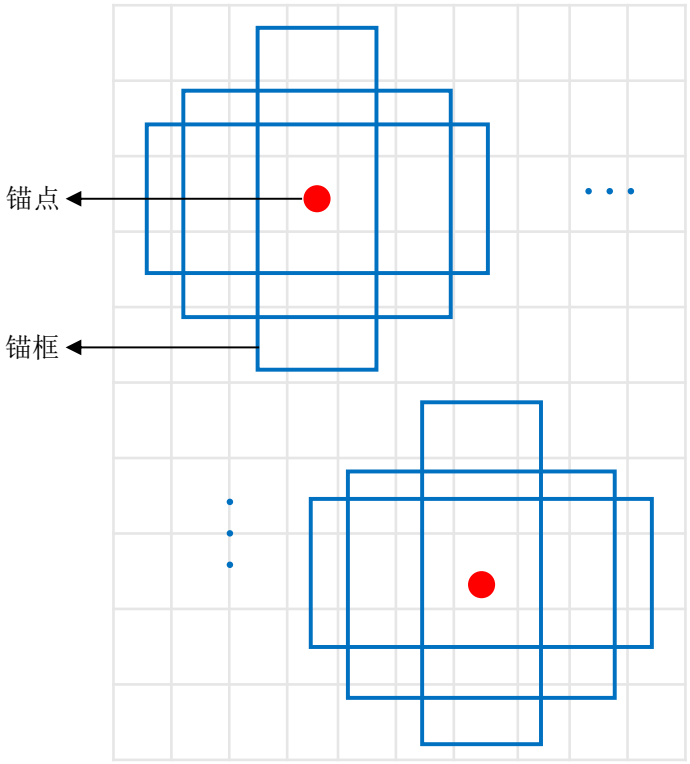


图 2-7 锚框示意图

Fig.2-7 Illustration of anchor box

Anchor-based 检测算法又可以被分为两级检测识别算法和单级检测识别算法。两级检测识别算法主要有 R-CNN[4]、Fast R-CNN[5]和 SPPNet[54]等，其中 SPPNet 在最后一个卷积层后接入了空间金字塔池化层，使用这种方式，可以让网络输入任意的图片，突破了 CNN 需要输入固定尺寸图像的限制。单级检测识别算法将分类任务和定位任务融合在一起完成，经过单次检测即可直接得到最终的检测结果，因此其速度一般比两级检测识别算法快，但精度有一些损失。常用的单阶段检测

器有 SSD、RetinaNet[55]和 RFB-Net[56]，其中 RetinaNet 提出了一种平衡样本的损失函数 Focal Loss 用于图像领域解决数据不平衡造成的模型性能问题。

对于 Anchor-free 目标检测算法，其摒弃了利用锚框去定位目标的方式。目前 Anchor-free 目标检测算法主要分为如下两类表示边界框的方法：

（1）基于关键点的检测算法：首先检测物体（如车辆部件）左上和右下的角点，然后连接角点组成了矩形边界框。

（2）基于中心的检测算法：直接检测目标的中心区域和边界信息，将分类和回归解耦为两个子网格。

2.3.3 数据增强

数据增强是对训练集进行各种变换，使训练集更加丰富，从而模型更加具有泛化能力，避免过拟合。在训练数据理想的情况下，模型可能会在遇到一些特殊情况时出现错误，如遮挡、亮度、模糊等。为了提高模型的鲁棒性并降低对图像的敏感度，可以对训练数据进行加噪、掩码等处理。此外，车辆部件通常是小目标样本，可以通过增加小目标样本的数量和类别来提高对小目标物体的检测性能。

图 2-8 Mosaic 数据增强

Fig.2-8 Mosaic data augmentation

数据增强较为常用的几何变换方法有翻转、旋转、裁剪、缩放和抖动等；较常用的像素变换方法有：加椒盐噪声、进行高斯模糊、调整 HSV 对比度、直方图均衡化和调整白平衡等。Cutout[57]和随机擦除[58]均考虑了物体被遮挡的特殊情况，提高了检测器的鲁棒性。对于小目标样本，YOLOv4 提出了 Mosaic 数据增强，将 4 张图片通过随机缩放、裁剪和排布的方式进行拼接，以增加小目标样本量，并赋予数据集更多样化和复杂性。图 2-8 展示了 Mosaic 增强后的图片和去除复杂背景的对比图。同样是增加小目标样本数量，Kisantal M[59]提出了一种更加简单粗暴的方法，直接采用复制粘贴（Copy-Paste）小目标的方法来进行数据增强，对

于本来只有一个物体（如车转向灯）的小目标对象，首先进行复制，然后粘贴几次到图片里与原物体不重叠的任何位置来增多小尺寸物体的数量，从而使得模型给予小物体更多的关注，以此来提高小物体的检测精度。

2.3.4 非极大值抑制算法

对于一个检测目标如汽车，检测器常常会输出多个预测框，其中常包含大量冗余的检测框。本节只希望得到贴近真实目标的检测框，为了去除冗余的检测框，采用非极大值抑制[60]（Non-Maximum Suppression，NMS）。对于最终生成的候选框（假定有n个）：B_1~B_n；每个候选框对应的分类分数：S_1~S_n；预先设置的NMS阈值：N_t，NMS 的算法流程如下：

a：建立一个最优框的集合 D，初始化为空集。

b：遍历 B_1~B_n 并进行排序，选取分类分数最高的预测框 B_m，将 B_m 从集合 B 中去除加入到集合 D。

c：遍历 B_1~B_n，分别与预测框 B_m 计算交并比，如果高于 N_t，则认为此预测框与 B_m 重叠，将此框从集合 B 中去除。

d：回到 b 阶段进行迭代，直至集合 B 为空集，集合 D 即为所求。

2.3.5 特征金字塔网络

为了检测小目标样本，研究人员常采用图像金字塔[61],但是图像金字塔需要大量的内存和计算资源，于是 Lin[10]等人提出了特征金字塔网络（Feature Pyramid Networks，FPN）。目标检测网络提取特征时，居前的卷积层提取的特征语义信息比较少，但能够实现准确的目标定位；而居后的卷积层提取的特征语义信息比较丰富，但目标位置比较模糊。因此针对骨干网络提取的相邻特征，FPN 通过自底向上，自上向下和横向连接的网络连接改变来进行多尺度特征融合，同时根据不同特征层进行预测。这种方式在基本上不增加模型计算量的同时，为整体网络对于小物体的检测精度带来了较大提升。

FPN 的灵感来自于低层特征高分辨率和高层特征的高语义信息的融合再使用，达到了极佳的效果。如图 2-9 所示，其中 (a) 是图像金字塔网络，特征是在每个图像尺度上独立计算的，因此其计算量大，需要大量内存；(b) 是 Faster-RCNN 中的检测方式，只关注了最深一层特征图的信息从而对于小目标检测帮助有限；(c) 是 SPPNet 网络的空间金字塔结构，其获得的特征不够鲁棒，没有充分利用浅层特征；(d) 即 FPN 中的特征金字塔结构，这种自底向上、自上向下的网络结构将多个不同尺寸的特征图融合在一起，充分利用了浅层特征和深层特征，提高了目标检测的性能表现。近些年来，基于 FPN 的研究层出不穷，例如 PANet[62]、ASFF[63]、和 BiFPN[15]证明了双向融合的有效性,同时尝试了更复杂的双向融合。

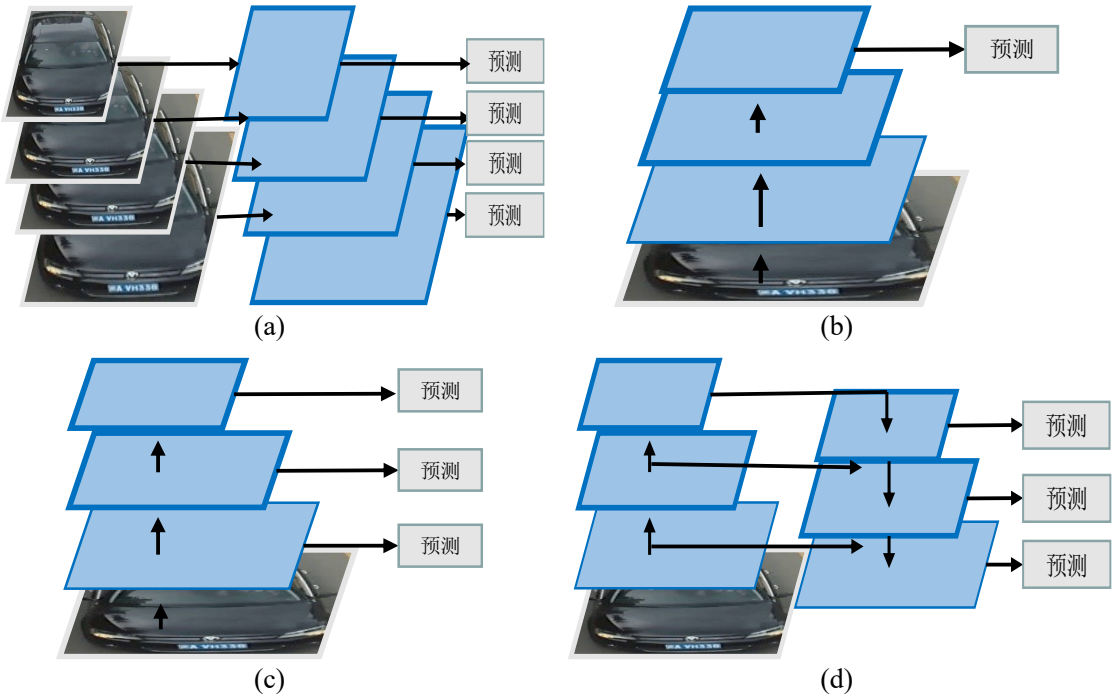


图 2-9 FPN 与其他网络结构的对比。 (a) 图像金字塔；(b) 单特征输出预测；(c) 空间金字塔层次结构；(d) 特征金字塔网络

Fig.2-9 FPN compared with other network structures. (a) Featurized image pyramid; (b) Single feature map; (c) Pyramidal feature hierarchy; (d) Feature Pyramid Network

2.4 目标检测器构成

如今，目标检测器通常由四个组件构成：Input、Backbone、Neck 以及 Head。

目标检测器需要获得经过数据预处理（如随机裁剪、归一化和调整尺寸）的图像作为输入，可以是一张或多张图片。Backbone 中文翻译为骨干网络，充当了整个目标检测网络的一部分，其作用是提取图像中的特征信息，故也被称为特征提取网络。在计算机领域中，通过不断卷积提取图像中的特征（例如颜色、纹理以及物体的空间关联等），同时缩小特征图尺寸以找到核心的部分。Neck 是目标检测网络中承上启下的关节，也被称为特征融合网络。这一部分通常插入在 Backbone 和 Head 之间来收集不同的特征图，然后将收集的特征进行融合，使得检测器学习到的特征更具多样性，有利于下一步 Head 的具体任务（如分类、回归和关键点），最终提高了整个检测网络的性能。Head 在目标检测网络中一般被叫分类定位网络，根据提取的特征来预测目标的位置和类别，主要作用是定位和分类。

2.5 车辆重识别概述

2.5.1 车辆重识别任务流程

车辆重识别是一种基于计算机视觉技术的智能检测技术，主要在一系列的复杂非重叠的场景下筛选识别出不同摄像头捕获的同一车辆图像。现如今车辆重识别大多是基于图片进行车辆的检索，它的数据集中包括训练集和测试集，训练集负责模型的训练和学习；测试集可以被分为查询集和检索集。 车辆重识别任务的基本流程如图 2-10 所示，首先将查询集和检索集所有包含的图像输入模型，并进行特征提取，其次将查询集的每一张车辆图片与检索集的每一辆车进行比对，然后以事先决定的距离度量方式（如欧氏距离）来计算两张图片的特征距离，距离越小代表相似度越高也代表与目标车辆更加匹配，最后按照特征间距由小到大进行排序得到车辆重识别的检索排列结果。

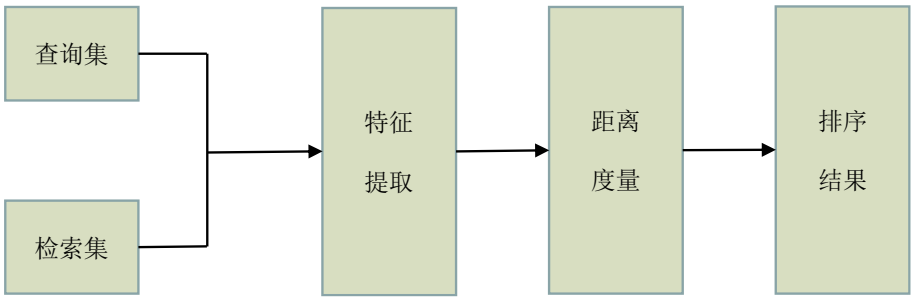


图 2-10 车辆重识别主要流程

Fig.2-10 Main process of vehicle re-identification

2.5.2 距离度量

距离度量是衡量两个图片的相似度大小，在人脸识别和车辆重识别等任务中起着决定性的作用，不同的任务要选择恰当的距离度量方式来匹配模型。车辆重识别是一个复杂的图像检索问题，充分考虑其距离度量的选择对于车辆重识别模型准确率提升有较大的帮助。在深度学习中，常见的距离度量方式有欧氏距离[64]、余弦距离、马氏距离和 KISSME[65]等，相关介绍如下：

（1）欧式距离

欧式距离，也称为欧拉距离，表征在欧式空间里两点之间的距离。欧氏距离简单通俗易于计算，但是它平等对待每个变量，没有考虑变量之间的差异。对于 N 维欧式空间中两点 x_1 和 x_2 之间的距离如下：

$$d=\sqrt{\sum_{i=1}^{N}(x_{1i}-x_{2i})^2} \tag{2-6}$$

（2）余弦距离

余弦距离是计算两个向量之间夹角余弦值的距离度量方法。与欧式距离不同，余弦距离更注重两个向量在方向上的差异，而非距离或长度上的差异。对于向量 A 和 B，N 维空间中的余弦距离公式如下：

$$cos\,\theta = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}} \tag{2-7}$$

式中，θ 为向量 A 和 B 方向的夹角。

（3）马氏距离

马氏距离是用于表示数据协方差距离的一种距离度量方法。与欧氏距离相比，马氏距离考虑了不同维度之间的相关性，因此更适合处理各个特征之间存在相关关系的数据。对于向量 x 和 y，马氏距离公式如下：

$$d_{(x,\,y)} = \sqrt{\left(x-\mu_x\right)^T \Sigma^{-1} \left(y-\mu_y\right)} \tag{2-8}$$

式中，μ_x、μ_y 分别为向量 x 和 y 的均值；Σ 为 x 与 y 的协方差。

（4）KISSME

从统计学的角度出发，Kostinger 等人[65]提出了 KISSME 距离度量算法。KISSME 主要利用假设检验的思想，以正负样本对概率的似然比函数为统计量，使得正负样本服从两个不同参数的高斯分布，学习一个马氏距离的协方差矩阵。本质上来讲，KISSME 是马氏距离的优化版。

首先，对于样本对 (x_i, y_j) 通过似然比观测它们之间的差异程度，构造似然比函数公式如下：

$$f(x_i, y_j) = log\frac{p(x_i - y_j|H_0)}{p(x_i - y_j|H_1)} \tag{2-9}$$

式中，H_0:(x_i, y_j) 表示样本对不相似（负样本）；H_1:(x_i, y_j) 表示样本对相似（正样本）。$f(x_i, y_j)$ 值与样本差异成正比，$f(x_i, y_j)$ 值越大，样本差异越大，反之，$f(x_i, y_j)$ 值越小，样本差异越小。

假设样本差向量满足 0 均值的高斯分布，则有：

$$f(x_i, y_j) = log\frac{\frac{1}{\sqrt{2\Pi|\Sigma_E|}} exp\left(-\frac{1}{2\left(x_i - y_j\right)^T \Sigma_E^{-1}\left(x_i - y_j\right)}\right)}{\frac{1}{\sqrt{2\Pi|\Sigma_I|}} exp\left(-\frac{1}{2\left(x_i - y_j\right)^T \Sigma_I^{-1}\left(x_i - y_j\right)}\right)} \tag{2-10}$$

式中，Σ_E 和 Σ_I 分别为不相似样本对和相似样本对的协方差。

将对数运算展开则有：

$$f(x_i\text{-}y_j)=\left(x_i\text{-}y_j\right)^T\Sigma_I^{-1}\left(x_i\text{-}y_j\right)+log\left(\left|\Sigma_I^{-1}\right|\right)$$
$$\text{-}\left(x_i\text{-}y_j\right)^T\Sigma_E^{-1}\left(x_i\text{-}y_j\right)\text{-}log\left(\left|\Sigma_E^{-1}\right|\right) \tag{2-11}$$

去掉只提供偏置的常数项，然后进行化简，最后得到反映对数似然比检验属性的马氏距离度量，表示如下：

$$d(x_i, y_j)=\left(x_i\text{-}y_j\right)^T\left(\Sigma_I^{-1}\text{-}\Sigma_E^{-1}\right)\left(x_i\text{-}y_j\right) \tag{2-12}$$

2.6 损失函数

损失函数是神经网络在训练时必不可少的一部分，其用于衡量模型的预测值和真实值不一样的程度。模型通过反向传播去更新参数，来降低真实值与预测值之间的损失，如何来优化损失函数对于模型的效率提升有着举足轻重的地位。车辆部件检测和重识别都是多任务学习：车辆部件检测中主要构建分类损失和回归损失两个子任务；车辆重识别中主要构建分类损失和以三元组损失为主的特征损失[66]。下面对三类损失中常用的损失函数进行介绍：

（1）分类损失

a）交叉熵损失函数

交叉熵[67]是为解决分类问题而提出的，对于二分类问题，需要预测的结果只有两种情况：对于两个类别预测得到的概率为 p 和 $1\text{-}p$，此时交叉熵损失如下：

$$L=\frac{1}{N}\sum_i L_i=\frac{1}{N}\sum_i\text{-}\left[y_i\cdot log\left(p_i\right)+\left(1\text{-}y_i\right)\cdot log\left(1\text{-}p_i\right)\right] \tag{2-13}$$

式中，y_i 表示样本 i 的标签；p_i 表示样本 i 预测为正类的概率。

在式 (2-13) 中，当且仅当 y_i 等于 p_i 时，交叉熵损失为 0，否则就是一个正数；当 y_i 和 p_i 相差越大，交叉熵损失就越大。交叉熵损失也可拓展到多分类情境下，如式 (2-14)，在其中 M 是待分类类别的数量。

$$L=\frac{1}{N}\sum_i L_i=\text{-}\frac{1}{N}\sum_i\sum_{c=1}^{M}y_{ic}\,log\left(p_{ic}\right) \tag{2-14}$$

b）Focal Loss

在单级检测识别网络中，由于没有提取候选框的步骤极易出现正负样本不平衡问题。Focal Loss[55]的提出解决了这个问题，其是基于二分类交叉熵损失的。Focal Loss 引入了一个调节因子来快速聚焦在那些难区分的样本，降低易分样本的损失贡献。公式如下：

$$FL=\begin{cases} -\alpha(1-p)^{\gamma}\log(p), & if\ y=1 \\ -(1-\alpha)p^{\gamma}\log(1-p), & if\ y=0 \end{cases} \tag{2-15}$$

式中，$(1-p)^{\gamma}$ 是调节因子；γ 是可调节的聚焦参数。

（2）回归损失

a）Smooth-L1 回归损失函数

L1 损失又称最小绝对值误差，最小化目标值 y_i 和估计值 $f(x_i)$的绝对差值之和：

$$L1=\sum_{i=1}^{n}\left|y_i-f(x_i)\right| \tag{2-16}$$

L2 损失又称为最小平方误差，其函数曲线连续，处处可导。相较于 L1 损失更加稳定，但是由于采用平方运算会放大误差，受离群点的影响比较大。L2 损失是最小化目标值 y_i 和估计值 $f(x_i)$的差值平方和：

$$L2=\sum_{i=1}^{n}\left(y_i-f(x_i)\right)^2 \tag{2-17}$$

L1 和 L2 损失相互组合就是 Smooth-L1 回归损失：

$$L_s=\begin{cases} 0.5x^2, & |x|<1 \\ |x|-0.5, & x<-1\ or\ x>1 \end{cases} \tag{2-18}$$

Smooth-L1 回归损失结合了 L1 和 L2 损失的优点，当 $|x|$ 大于 1 时，相当于 L1 损失，解决了梯度较大时加大误差的问题；当 $|x|$ 小于 1 时，相当于 L2 损失，解决了难以收敛的问题。Smooth-L1 回归损失既鲁棒又稳定且有单一解。

b）IOU 损失

IOU 损失[68]是广泛使用的回归损失函数，其具有尺度不变性。它把真实框和预测框的四个坐标点关联起来，使用两者相交区域和合并区域面积的比值，也就是交并比作为损失函数。对于真实框 A 和预测框 B，IOU 损失公式如下：

$$L_{IOU}=-\ln\frac{|A\cap B|}{|A\cup B|} \tag{2-19}$$

IOU 损失并不能完全反映重合的好坏，也无法比较预测框和真实框的远近。基于此，研究人员基于 IOU 损失进行了一些改进，如 GIOU[69]，DIOU[70]等

（3）三元组损失函数

三元组损失函数[71]（Triplet Loss）最早是由 Google 在人脸识别任务中提出来的，其目的是做到非同类相似样本的区分。Triplet Loss 的优势是细节划分，输入相似时，可以学到输入更好的隐层表示。在数据集中随机选取原样本（Anchor），与原样本同类的为正样本（Positive），不同类的为负样本（Negative），这样构成

了一个三元组（原样本，正样本，负样本）。如图 2-11 所示，对于三元组，Triplet Loss 会使得原样本特征更加靠近正样本特征，同时更加远离负样本特征。

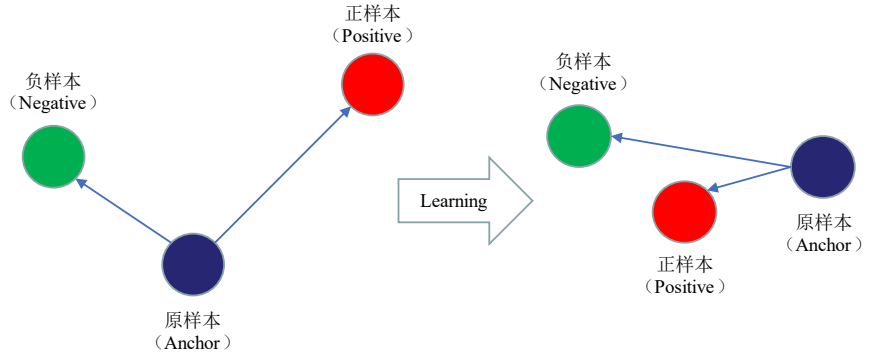


图 2-11 三元组损失学习示意图

Fig.2-11 Illustration of triplet loss learning

对于一个样本容量为 N 的三元组 (x^a, x^p, x^n)，Triplet Loss 表示如下：

$$L_{triplet}=\sum_i^N\left[\left\|f(x_i^a)-f(x_i^p)\right\|_2^2-\left\|f(x_i^a)-f(x_i^n)\right\|_2^2+\alpha\right]_+ \tag{2-20}$$

式中，$f(x)$是 x 的特征表示；i 为样本的序列；α 是预设的间隔量为保证训练朝着预设目标进行。

2.7 评价指标

任何模型都需要一个衡量指标评价其好坏，对车辆部件检测模型来说，衡量指标主要有精确率（Precision，P）、平均精度（Average Precision，AP）、平均精度均值（mean Average Precision，mAP）和浮点运算数（Floating Point Operations，FLOPs）；对重识别模型来说，衡量指标主要有 mAP 和 Rank-k（k 一般取 1、5、10 等）。

（1）车辆部件检测的评价指标

针对车辆部件的某一类（记为 A 类）数据样本图片，A 类区域为正样本，其余背景为负样本，只有当预测框分类正确且 IOU 大于预定义的阈值时，将其视为正确样本，否则为错误样本。具体会产生四种不同的样本：真的正样本（True Positive，TP）：检测到 A 类预测框且被正确分为 A 类；真的负样本（True Negative，TN）：未检测到 A 类以及其他预测框；假的正样本（False Positive，FP）：背景检测到 A 类预测框被错误分 A 类；假的负样本（False Negative，FN）：未检测到 A 类预测框但存在 A 类。由此可得到模型的精确率和召回率为：

$$P=\frac{TP}{TP+FP} \tag{2-21}$$

$$R=\frac{TP}{TP+FN} \tag{2-22}$$

由精确率和召回率就可以得到 A 类物体的平均精度：

$$AP=\int_0^1 P(R)dR \tag{2-23}$$

对于多类别的目标检测，mAP 即是每类的 AP 做均值处理。mAP 是对模型精度上的衡量，而模型参数量和 FLOPs 是在模型效率上的衡量。模型参数量是网络结构需要训练的参数总量，可以用来评判模型的大小以及所占内存；FLOPs 是浮点运算数，可以用来评判模型或算法的复杂度。

（2）车辆重识别的评价指标

车辆重识别任务中也有 mAP 衡量指标，但是与车辆部件检测有一点不同，即 AP 的计算方式如下：

$$AP=\frac{\sum_{k=1}^{n} P(k)\times gt(k)}{N_{gt}} \tag{2-24}$$

式中，n 是检索集中图像总数；N_{gt} 是目标车辆的总样本数，$P(k)$是返回排列结果前 k 图像的检索精度，若第 k 位置匹配正确，则 $gt(k)=1$，否则 $gt(k)=0$。

mAP 则是对于查询集的目标图片的 AP 做均值处理：

$$mAP=\frac{\sum_{i=1}^{n} AP(i)}{n} \tag{2-25}$$

Rank-k 又称为前 k 击中率，常用于车辆重识别任务。如 2.5.1 节所述，对于每张目标车辆图片会得到特征距离由小到大的排列，Rank-k 是排列中前 k 张车辆图片中有正确结果的概率。其公式表示如下：

$$Rank\text{-}k=\frac{\sum_{i=1}^{n} S_{i,k}}{n} \tag{2-26}$$

其中 n 为待检索的图片的总数，对于第 i 个检索图片，车辆重识别的检索结果中前 k 个有正确匹配图片时，$S_{i,k}=1$，否则 $S_{i,k}=0$。

2.8 本章小结

本章首先系统的阐述了卷积神经网络的基本构成，并简要的梳理了几种经典的网络模型，其次分别概述了车辆部件检测和车辆重识别的问题描述以及理论基础，主要介绍了目标检测器的主要构成、锚框、数据增强、非极大值抑制、特征金字塔网络以及车辆重识别的距离度量方式等，然后解释了车辆部件检测与重识别中分类损失、回归损失以及三元组损失的基本原理。为后续章节中判定模型的优劣做铺垫，最后介绍了在任务中用到的评价指标。

第 3 章 优化 EfficientDet 的车辆部件检测算法

基于上一章阐述的相关理论，本章提出了一种高效、高精度的车辆部件检测算法，主要阐述了 5 部分内容，分别为引言，EfficientDet 模型概述，改进的 EfficientDet 车辆部件检测算法，实验结果与分析和本章小结。在 3.1 节，简单分析了车辆部件检测任务，并引出本章的创新性研究工作；在 3.2 节，概述了 EfficientDet 基础模型；在 3.3 节，详细阐述了对于 EfficientDet 模型的改进与分析；在 3.4 节，对提出算法进行实验以证明该方法的有效性；在 3.5 节中，对本章内容进行总结。

3.1 引言

基于深度学习的目标检测方法取得了巨大的进展，卷积神经网络结构也在 2016 年被应用在车辆目标识别任务上，到现在为止具体到车辆部件检测的研究却很少。车辆部件检测属于目标检测的任务范围之内，车辆部件检测任务也可以使用目标检测算法。深度学习的手段依靠深度 CNN 逐层提取图像的高级特征，这往往表现的更精确、更有效。但是更精确的同时伴随着模型的大型尺寸和昂贵的计算成本，例如，文献[72]提出的 NAS-FPN 探测器需要 167M（模型参数量）和 3045B（FLOPs）达到最先进的精度，这比 RetinaNet 多 30 倍，阻碍了其在现实世界应用中的部署。更好的检测模型通常会牺牲一些准确性，并且只关注特定或较窄范围的资源需求，但不同的现实世界应用程序（例如车辆部件检测系统）通常需要不同的资源限制。

最关心的问题往往是在高精准度的同时兼顾模型的尺寸和计算成本以提高其效率，所以本章在 EfficientDet 的基础上构建了一个高效，高精度的无锚框车辆部件检测模型（EFDet-SPP）。该方法的创新性工作如下：

（1）建立车辆部件检测数据集，确保数据充足，场景复杂且具有普遍性，同时在输入网络时将 Mosaic 和 Copy-Paste 数据增强方式相结合来针对性增强车辆部件这类小目标。

（2）在骨干网络末尾连接空间金字塔池化模块，对特征图进行多尺度采样，串联特征且有效地捕获高语义信息。

（3）在特征融合网络（BiFPN）的基础上增加纵向交叉跨层连接以充分利用不同层级的语义信息和空间信息，实现多尺度特征的自适应融合，提升模型的精度和鲁棒性。

（4）基于锚框预测转变为基于像素点预测，消除了与锚框相关的超参数同时减少了计算交并比时计算量和内存占用，提高模型的泛化能力和在不同场景下的适应性。

3.2 EfficientDet 模型概述

2019 年，谷歌提出一个效率高的分类网络 EfficientNet[73]，是一种标准化模型扩展结构，其通过调整输入图像的大小、网络的深度与卷积通道数来实现网络在精度和效率上的平衡优化。而 EfficientDet 是以 EfficientNet 为骨干网络的基于锚框的单阶段目标检测算法，同时采用了多种优化，如使用 BiFPN 进行双向特征融合和采用复合缩放方法均匀缩放各个部分的分辨率、深度和宽度。这些优化使得 EfficientDet 能够在广泛的资源约束下获得更好的精度和效率。在相似的准确性约束下，EfficientDet 模型在 GPU 上比其他检测器快 2-4 倍，在 CPU 上快 5-11 倍。此外，EfficientDet-D7 在 COCO test-dev 测试集上 mAP 达到 55.1%，与之前表现优秀的目标检测模型相比，使用的参数减少了 4-9 倍，同时，FLOPs 减少了 13-24 倍。

EfficientDet 的网络结构如图 3-1 所示，从图中可以看出其网络结构主要包括特征提取网络，特征融合网络和目标分类和定位的预测网络。

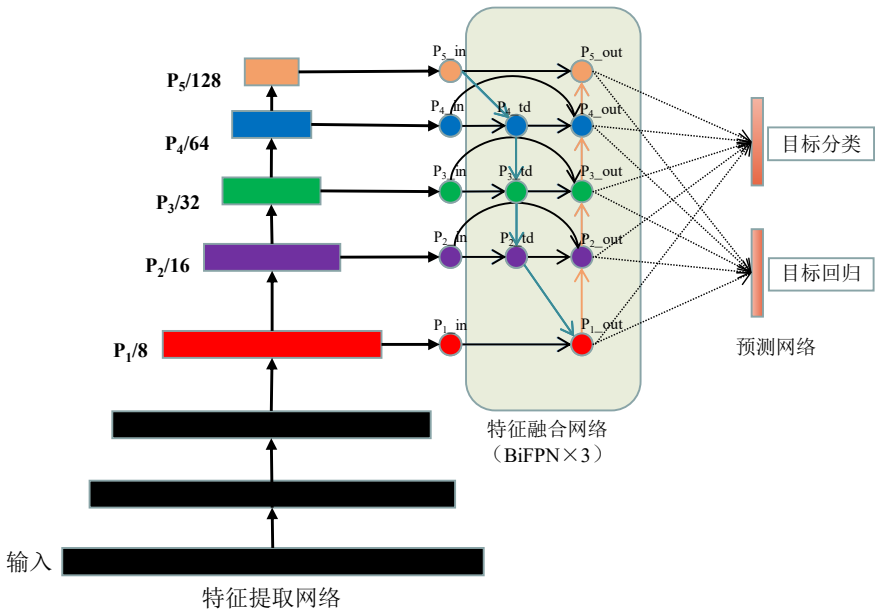


图 3-1 EfficientDet 网络架构图

Fig.3-1 EfficientDet architecture

特征提取网络直接使用的是 EfficientNet，其能达到较高的准确度，同时不需要大量的算力资源。以 EfficientNet-B0 为基准，网络主体部分主要分为两部分：首层和尾层使用的卷积核为 3×3 的普通卷积和其余层使用的重复堆叠的 MBConv 卷积模块。MBConv 包含 5 个主要操作分别为 1×1 卷积进行升维：输入的特征图经 1×1 卷积，卷积核个数为输入特征图通道的 n（$1<n<6$）倍，从而将通道数升维；深度可分离卷积：包含深度卷积和逐点卷积两步操作，用于提取特征； SE 注意力机制模块（Squeeze-and-Excitation）：特征的重要性权重计算，从而增强重

要特征的响应；1×1 卷积进行降维：将维度调整为输出通道数；Dropout 层：随机舍弃用于减少过拟合。MBConv 的结构如图 3-2 所示。

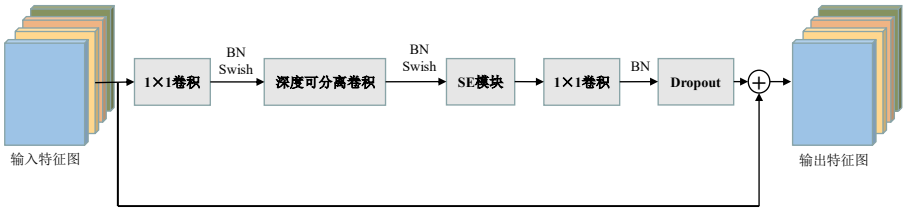


图 3-2 MBConv 结构

Fig.3-2 Structure of MBConv

图 3-2 可看出，MBConv 在升维 1×1 卷积和深度可分离卷积中都包含了 BN 层和 Swish 激活函数，并加入了 SE 模块。SE 模块由一个全局平均池化、两个全连接层和激活函数组成。首先，全局平均池化对输入的特征图进行压缩得到每个通道的平均值，然后通过第一个全连接层和 Swish 激活函数降低维度，再通过第二个全连接层和 Sigmoid 激活函数恢复维度，这样就得到每个通道的权重，最后，将权重乘以原始的特征图，实现通道间的注意力机制。这样，SE 模块可以动态地调整特征图中每个通道的重要性，从而提高网络的表达能力。SE 模块见图 3-3。

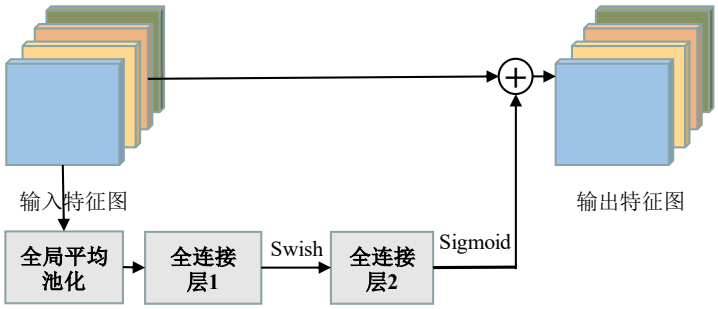


图 3-3 SE 模块结构

Fig.3-3 Structure of SE module

从图 3-1 可以看到 BiFPN 特征融合网络连接在 EfficientNet 主干网络之后，用于简单和快速的多尺度融合 5 个初步提取的特征图。BiFPN 在 FPN 的基础上通过有效的双向交叉连接（自顶向下和自底向上）和加权特征融合使网络更加关注重要的层次，而且还减少了一些不必要的层的结点连接。同时在横向连接增加了一条额外的边，将输入结点与输出结点直接相连，相当于一个残差连接，可以在不增加计算量的同时融合更加复杂的特征。

多尺度特征融合目的在于聚合不同尺寸的特征。与以往 FPN 调整成统一尺寸再直接相加的方式不同的是，BiFPN 的加权特征融合是一种学习不同输入特征的重要性，针对不同的输入特征采用不同的融合方式。具体来说，对每个层级的特

征增加了一个可学习的权重向量，这样可以保证特征质量，避免信息冗余。EfficientDet 设计了三种加权特征融合的方法如下：

$$O=\sum_i \omega_i \cdot I_i \tag{3-1}$$

$$O=\sum_i \frac{e^{\omega_j}}{\sum_j e^{\omega_j}} \cdot I_i \tag{3-2}$$

$$O=\sum_i \frac{\omega_i}{\epsilon+\sum_j \omega_j} \cdot I_i \tag{3-3}$$

式中，w_i为可学习的权重，对于特征来说是一个标量；ϵ 为0.0001，避免数据不稳定。式 (3-3) 即快速归一化融合有稳定约束且计算量小。最终将融合后的特征图输入目标分类和定位的预测网络输出最终结果。EfficientDet 的检测精度较与其他目标检测网络有明显的优势，实现了速度与精度的均衡。

3.3 改进的 EfficientDet 车辆部件检测算法

本章提出了改进 EfficientDet 的车辆部件检测模型（EFDet-SPP），网络结构如图 3-4 所示。在车辆部件检测场景下，对输入、BackBone、Neck 和 Head 进行了针对性的改进与优化。

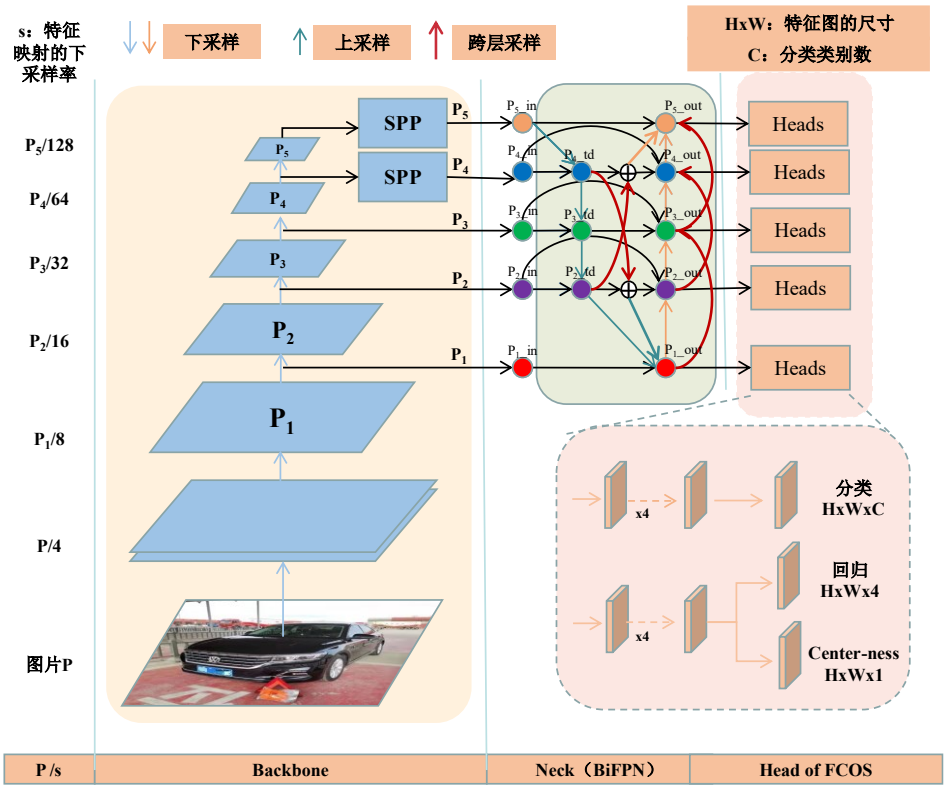


图 3-4 车辆部件检测方法（EFDet-SPP）网络结构图

Fig.3-4 Vehicle component detection method (EFDet-SPP) network structure

3.3.1 图像数据增强策略

目标检测方面，现在有很多普遍性数据集：VOC 数据集[74]、COCO 数据集[75]；车辆检测方面，也有不少车辆检测数据集：BDD100K[76]、CompCars[77]。但是这些数据集在特定的场景下（车辆部件，精细识别）并不适用，故笔者采集了不同地区车管所，检测站以及道路监控捕获的大量车辆及其部件图像，根据在一张图片上目标的密集程度和占比分为了车辆大部件数据集（Vehicle Large Component dataset，VLC）和车辆密集部件数据集（Vehicle Dense Component dataset，VDC）。首先针对数据集中每一类型的标签统计观察，如表 3-1 所示：

表 3-1 数据集大、中、小标签框统计

Table3-1 Statistical of dataset large, medium and small labels

数据集	$area<32^2$	$32^2 \le area \le 96^2$	$area>96^2$	AP_{small}	mAP
VLC	1154	11562	48114	0.563	0.846
VDC	108579	82314	32141	0.435	0.789

注：AP_{small} 和 mAP 是 EfficientDet-D0 模型测试；mAP 均是在 IOU=0.5 的条件下测试。

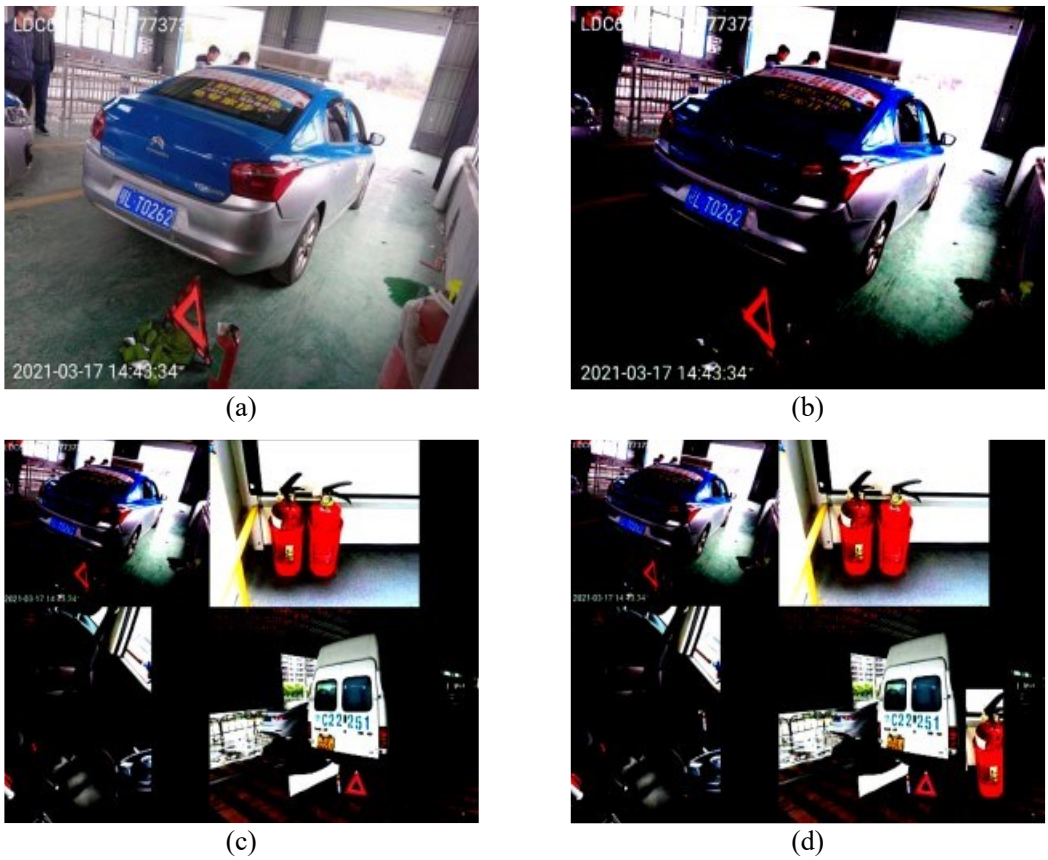
由表 3-1 可知，EfficientDet 模型在车辆部件检测场景下表现不佳，VLC 数据集的小目标样本数量较少，而且在两个数据集上均对小目标样本的识别不佳，导致 AP_{small} 很小，无法进行有效的模型训练。因此，首先将图像数据进行了去均值中心化，公式如下：

$$x_1=\frac{x-\mu}{255\sigma} \tag{3-4}$$

式中，μ 和 σ 分别代表均值和方差。之后结合了 Mosaic 和复制粘贴两种数据增强方式来对输入图像进行处理，Mosaic 是将四张图像拼接成一张新的图像，并保留每张图像对应的标注框，复制粘贴是将一张图像中的目标对象复制到其他位置，并更新标注框。本节在复制粘贴过程中添加附加条件使得其只针对小目标生效，这些方法的好处是可以丰富数据集中的场景和对象组合，增加模型的泛化能力和鲁棒性。

本节对于图像输入 EFDet-SPP 网络之前进行的数据增强处理过程进行了可视化如图 3-5 所示。对于一张 512×512 的三通道彩色图像，首先进行去均值中心化，然后将处理后的图像与其他随机裁剪后的三张图像进行拼接成为一张新的图像，最终对图像对应的标签进行筛选，针对性的将小目标样本标签复制然后粘贴到与原标签区域不重合的随机区域，这样就得到 EFDet-SPP 的输入图像。在现实生活

中的应用场景下，数据复杂多变，数据增强可以使数据样本更加复杂，从而提高目标检测模型的鲁棒性。



(a) (b)

(c) (d)

图 3-5 数据增强的可视化。 (a) 原图像；(b) 去均值中心化图像；(c) Mosaic 图像；(d) 复制粘贴图像

Fig.3-5 Data augmentation visualization. (a) Original image; (b) Regularization image; (c) Mosaic image; (d) Copy-pasty image

3.3.2 空间金字塔池化

在目标检测领域，通常使用卷积和池化来提取并且缩小特征图。为了能够得到一个高效的车辆部件检测网络，本章选择了较低的图像分辨率，但这样会降低每个像素中包含的语义信息，特别是对于小尺寸车辆部件。因此，为了提高其检测效果，引入了空间金字塔池化模块（Spatial Pyramid Pooling，SPP）。SPP 模块是一种用于处理不同输入尺寸的图像的池化方法，在空间上分成了不同大小（$m \times m$）的最大池化层，其中 m 可以分别为 1、3、5 等值。针对每个特征图用不同大小的最大池化层处理，拼接起来形成一个固定大小的特征向量，然后可以进行进一步的融合处理。这样可以增大感受野，多尺度提取丰富的特征，在一个单一

的过程中形成整个图像的特征映射，提高车辆部件检测的精度和鲁棒性。特征提取网络在卷积过程中的卷积核大小、通道数等细节信息列于表 3-2。

表 3-2 EFDet-SPP 特征提取网络结构

Table3-2 The structure of EFDet-SPP backbone network

阶段	网络层	特征图（H×W）	通道数
1	Conv3×3	512×512	32
2	MBConv1，k3×3	256×256	16
3	MBConv6，k3×3	128×128	24
4	MBConv6，k5×5	64×64	40
5	MBConv6，k3×3	32×32	80
6	MBConv6，k5×5	16×16	112
7	MBConv6，k5×5	8×8	192
8	MBConv6，k3×3	4×4	320

注：*k* 表示的是卷积核大小。

由表 3-2 可知，EFDet-SPP 的输入图像尺寸为 512×512，通过一个 3×3 的普通卷积和重复堆叠的 MBConv 卷积会得到不同尺寸的五个特征图，具体表示如下：

$$P_1: 64\times64,\ P_2: 32\times32,\ P_3: 16\times16,\ P_4: 8\times8,\ P_5: 4\times4 \tag{3-5}$$

式中，P_1~P_5 为最终输出的五个特征图。高层特征图蕴含了较多的语义信息和全局信息，而低层特征图包含了更多的细节信息和局部信息。为了多尺度提取更丰富和鲁棒的特征，于提取的 P_4 和 P_5 特征图后加入了不同尺度的 SPP 模块如图 3-6 所示。

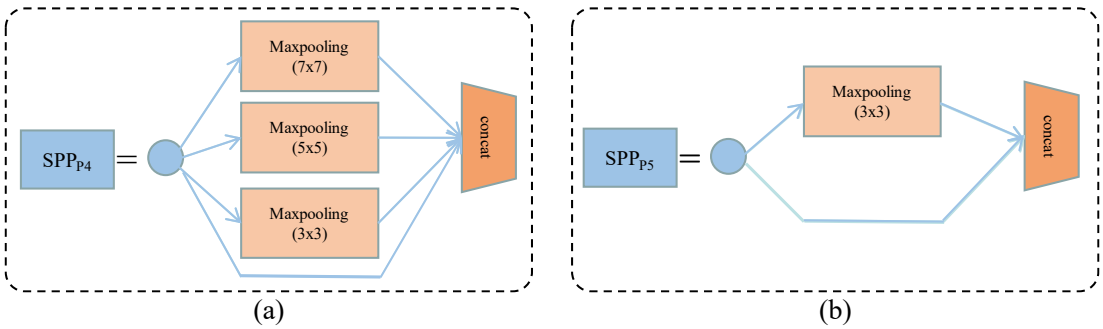


图 3-6 不同的空间金字塔池化模块。(a) 对于 P_4 的 SPP 模块；(b) 对于 P_5 的 SPP 模块

Fig.3-6 Different Spatial Pyramid Pooling modules. (a) SPP module for P_4; (b) SPP module for P_5

3.3.3 纵向交叉跨层连接的 BiFPN

在 BiFPN 中的横向数据流中，作者添加了跨层连接，这样可以缩短上下层之间信息传递的路径，实现更高层次的特征融合。在本章的设计中，起初的优化方向是如何来充分利用纵向跨层连接更多的结点，从而融合得到更加丰富的特征信息。所以设计了如图 3-7 所示的网络结构。

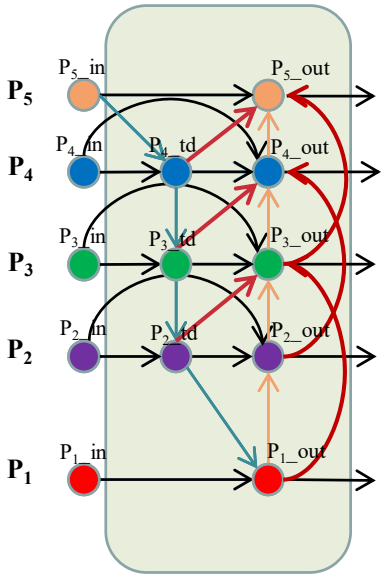


图 3-7 BiFPN 初始改进方案

Fig.3-7 BiFPN initial improvement programme

受到 BiFPN 中横向跨层连接以及残差网络的思想启发，本章采用了一种新颖的路径融合方式。添加了大量的纵向跨层连接，使得不同层级之间的信息传播更加快速和有效。相比其他连接方式，跨层连接在反向传播时有更短的路径。在初始方案里，增加了从低层到高层的跨层连接，在中间和两端都是相邻层级节点的信息流。在跨多少层特征结点的问题上，最初的想法是如果融合特征节点之间的层级差距太大，那么它们所融合的底层信息会过多。这样的话在特征融合时，底层信息如果学习到的权重过大，那么就打破了位置信息和高层结点原有的细节信息的平衡。那样对网络的精确性和鲁棒性存在消极作用，所以最终决定添加的跨层连接都只跳一层。希望可以让高层结点尽可能的融合更多底层位置信息，有利于对后面车辆小部件的检测。

但是实验效果并不理想，这种设计方案在本文的数据集 VDC 上，始终要比原始的 BiFPN 的 mAP 精度要低 0.9%，没有提高检测效果。经过仔细分析，原因为：其一是过分强调了底层到顶层的信息传递，而忽视了底层结点；其二是顶层结点和底层结点的输入流差异较大，大多把底层特征融合到上层结点一起学习，如 P_{4_out} 和 P_{5_out} 等顶层结点分别有 5 条和 4 条输入流，而 P_{1_out} 和 P_{2_out} 结点则分别有

2 条和 3 条输入流。这样会导致高低层之间信息分配不均衡，一旦权重失衡，则会出现极端情况。

基于对初始改进方案的问题分析，最终本节设计了纵向交叉跨层连接的 BiFPN 如图 3-8(b) 所示。

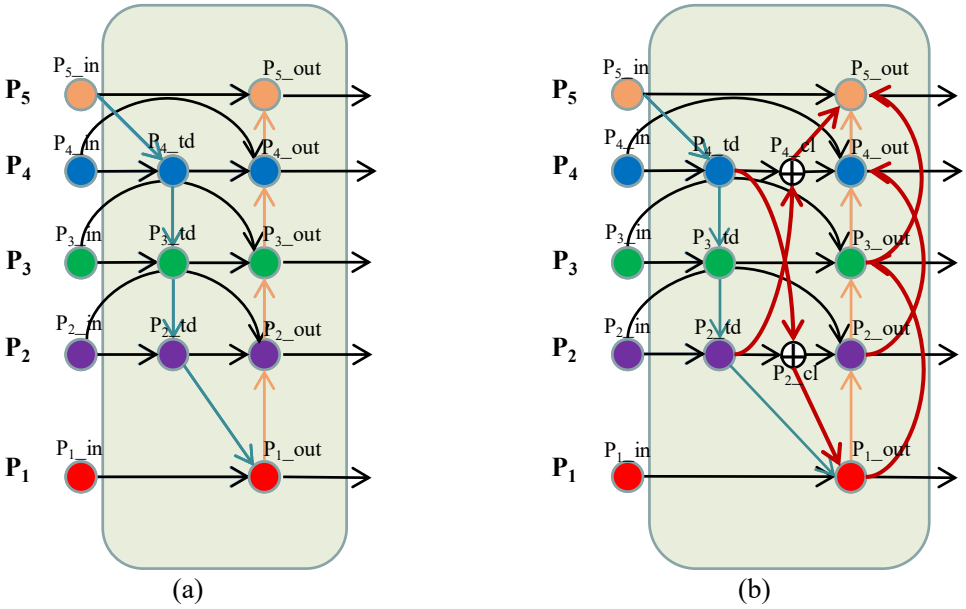


图 3-8 纵向交叉跨层连接的 BiFPN 网络结构对比。 (a) 原始 BiFPN；(b) 纵向交叉跨层连接的 BiFPN

Fig.3-8 Comparison of BiFPN network structures with longitudinal cross-layer connection. (a) Original BiFPN; (b) BiFPN with longitudinal cross-layer connection

在图 3-8 中，(a) 是原始的 BiFPN，它结合了 FPN 中的特征融合思想，使数据可以自顶向下和自底向上两个方向流动，同时采用了快速归一化融合方式进行特征融合。这里以 P_4 层为例，介绍特征融合的输出表达式：

$$P_4^{td}=Conv\left(\frac{\omega_1\cdot P_4^{in}+\omega_2\cdot Resize(P_5^{in})}{\omega_1+\omega_2+\epsilon}\right) \tag{3-6}$$

$$P_4^{out}=Conv\left(\frac{\omega_1'\cdot P_4^{in}+\omega_2'\cdot P_4^{td}+\omega_3'\cdot Resize(P_3^{out})}{\omega_1'+\omega_2'+\omega_3'+\epsilon}\right) \tag{3-7}$$

式中，P_4^{in} 和 P_5^{in} 是第四层和第五层的输入特征；P_4^{td} 是第四层自顶向下路径的中间特征；P_3^{out} 和 P_4^{out} 是第三层和第四层自底向上路径上的输出特征；ω_1 和 ω_2 是 P_4^{td} 的输入权重；ω_1'、ω_2' 和 ω_3' 是 P_4^{out} 的输入权重。

图 3-8(b) 是纵向交叉连接的 BiFPN，其中红色线条为增加的跨层数据流。本节在初始方案网络设计的仔细分析的基础上，采用了新的跨层连接策略，即纵向

交叉连接。这种方式综合平衡了高层结点和底层结点的输入数据流，同时充分地利用了不同层级的语义信息和位置信息。在第 2 层和第 4 层添加了两个交叉结点，缩短了底层特征图到顶层特征图的距离，更好的融合了底层和顶层特征信息，实现了更好的信息交互。这种设计也不会使结点的特征信息占比发生明显改变，最终得到了纵向交叉跨层连接的 BiFPN。改进后其特征融合方式也有所区别，只有自顶向下路径的中间特征融合没有改变，仍以 P_4 层为例，改进后的输出表达式表示如下：

$$P_4^{cl}=Conv\left(\frac{\omega_1\cdot P_4^{td}+\omega_2\cdot Resize(Resize(P_2^{td}))}{\omega_1+\omega_2+\epsilon}\right) \tag{3-8}$$

$$P_4^{out}=Conv\left(\frac{\omega_1'\cdot P_4^{in}+\omega_2'\cdot P_4^{cl}+\omega_3'\cdot Resize(P_3^{out})+\omega_4'\cdot Resize(Resize(P_2^{out}))}{\omega_1'+\omega_2'+\omega_3'+\omega_4'+\epsilon}\right) \tag{3-9}$$

3.3.4 无锚框预测方式

EfficientDet 是基于锚框的单阶段目标检测算法，它的检测性能在一定程度上受人工预先设定的锚框影响。在 2.3.2 节，简单的介绍了锚框的形成，在锚框的形成过程中有三个需要人工设置的超参数：尺度、长宽比以及数量。在不同的应用场景下，如车辆部件检测任务中，锚框的超参数就需要根据物体类别来人工动态调整。

由表 3-1 可看出，EfficientDet 网络在车辆部件检测场景下表现不佳，只是切换不同的应用场景，精度则达不到理想效果。经过慎重的分析，原因在于其依赖于一套预定义的锚框，在不同的应用场景下，数据集也存在巨大的差异性，锚框的尺寸、宽高比和数量是否合适对检测性能有决定性作用。针对锚框的问题，有两种优化的方向：其一，采用聚类方法去寻找合适的超参数；其二，改变基于锚框的预测方式，放弃锚框机制，去除关于锚框预定义的超参数。但是为了将召回率提高，Anchor-based 检测网络会在图像上密集放置锚框，如对于输入图像尺寸大于 800×800 的 FPN 设置了 180000 个锚框。 为了确定锚框是否为正样本，还需要大量关于交并比的复杂计算，这也占用了较大的计算内存。同时为了规避与锚框相关的对检测性能敏感的预定义超参数，本节选择后者方案借鉴 FCOS 检测模型的无锚框思想对 EfficientDet 进行针对性的改进。

在图 3-4 中可以看出，EFDet-SPP 的检测头部是三分支的共享预测网络，包括分类、回归和中心点回归（Center-ness），三者共享特征融合网络输出的 5 个特征，逐像素预测的方式来定位和分类。与 EfficientDet 不同的是它直接对特征图中每个

位置 (x, y) 对应原图的边框进行回归，将每一个像素点 (x, y) 都作为训练样本。对于特征图上每个位置 (x, y)，与原图像的映射公式如下：

$$(x', y') = \left(\left\lfloor \frac{s}{2} \right\rfloor + xs, \left\lfloor \frac{s}{2} \right\rfloor + ys\right) \tag{3-10}$$

在训练过程中，如果位置 (x, y) 对应的像素点在真实边框内部，则为正样本并且将该位置的类标签设置为真实边框的类标签，否则，作为负样本参与训练。在对位置回归时，则是分别计算像素点与真实边框四边的距离，得到一个 4 维的用于回归的偏移向量 $\eta^* = (l^*, t^*, r^*, b^*)$，对于真实边界框的集合 $Bi = (x_0^{(i)}, y_0^{(i)}, x_1^{(i)}, y_1^{(i)}, c)$，回归的任务可以表示为：

$$l^* = x - x_0^{(i)},\ t^* = y - y_0^{(i)},\ r^* = x_1^{(i)} - x,\ b^* = y_1^{(i)} - y \tag{3-11}$$

式中，$(x_0^{(i)}, y_0^{(i)})$ 和 $(x_1^{(i)}, y_1^{(i)})$ 分别为真实框左上右下的角点坐标；如图 3-9 所示，l^*, t^*, r^*, b^* 表示该位置到边界框四条边的距离。

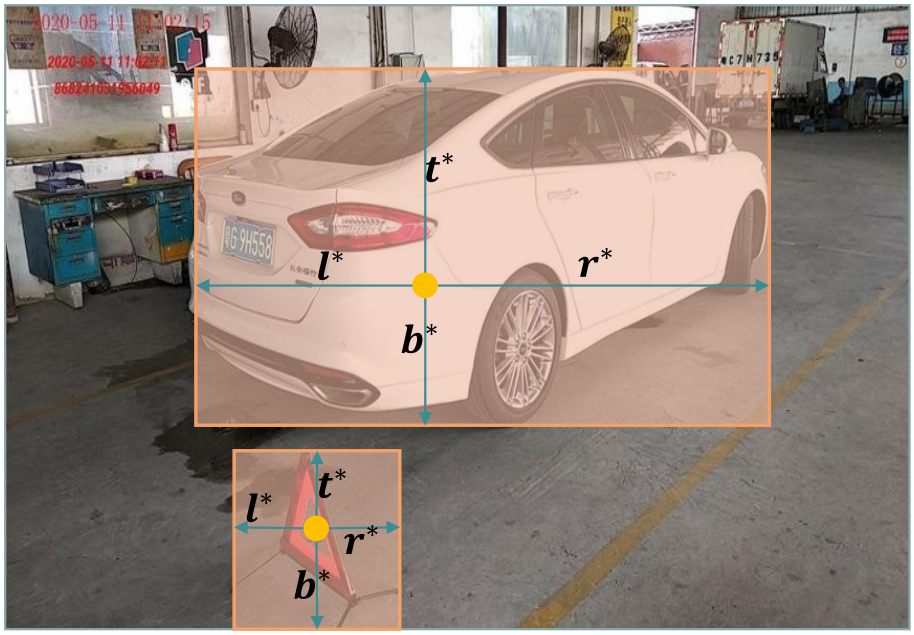


图 3-9 EFDet-SPP 预测的四维向量

Fig.3-9 Four-dimensional vector predicted by EFDet-SPP

另外一个重要的问题则是模糊样本，如图 3-10 所示，当一个位置 (x, y) 落在多个真实框内，此时 (x, y) 则为模糊样本。这时可以利用交叉跨层连接 BiFPN 的多级预测结构，选择最小的预测框作为回归目标。特征图中一个位置 (x, y) 满足式 $max(l^*, t^*, r^*, b^*) > m_i$ 或者 $max(l^*, t^*, r^*, b^*) < m_{i-1}$ 则该位置为负样本，其中 m_i 代表第 i 个特征层的最大回归距离。在本章中，将 m_0、m_1、m_2、m_3、m_4、m_5 分别设置

为 0、32、64、128、256、512。若有两个真实框都符合要求，则选择面积较小做回归预测。

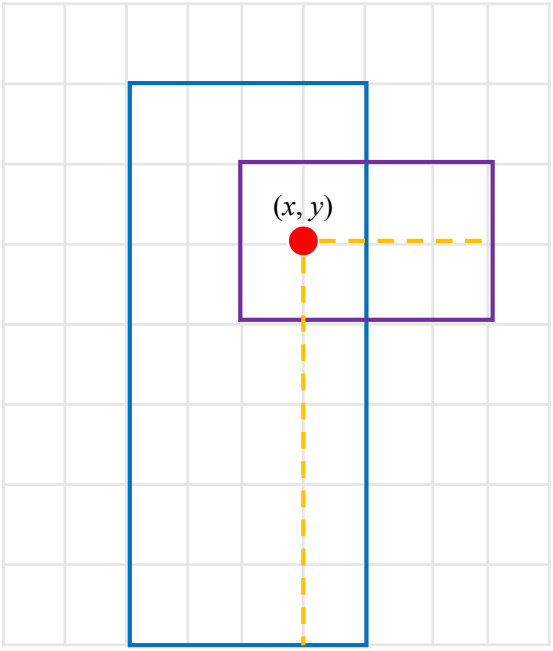


图 3-10 模糊样本示例图

Fig.3-10 An example of fuzzy sample

在图 3-10 中，假设位置 (x, y) 为 P_2 层特征，每一个网格代表 10 像素的话，距蓝色框的最大距离为 50，距紫色框的最大距离为 30，故选择蓝色框进行回归预测。若一个网格代表 12 像素的话，两个均在范围内，此时选择面积更小的紫色框进行回归预测。

基于像素级别的回归操作，在远离中心靠近真实框边缘的位置会产生大量的低质量预测框。为了去抑制低质量预测框，在回归分支旁引入了 Center-ness 分支。其计算公式如下：

$$Center\text{-}ness^{*}=\sqrt{\frac{min\left(l^{*}, r^{*}\right)}{max\left(l^{*}, r^{*}\right)}\times\frac{min\left(t^{*}, b^{*}\right)}{max\left(t^{*}, b^{*}\right)}} \tag{3-12}$$

从式中可以看出，位置处于真实框边界时，Center-ness 几乎为 0。将其与对应的类别置信度相乘计算最终得分，然后通过 NMS 算法进行过滤，得到最后的检测框。

损失函数在训练模型过程中具有至关重要的作用，它使模型逐渐收敛。不同的损失函数会对模型最终的收敛产生不同的效果，对于三个分支采用的损失函数表示如下：

$$L\left(\{p_{x,y}\},\{t_{x,y}\},\{s_{x,y}\}\right)=\frac{1}{N_{pos}}\sum_{x,y}L_{cls}\left(p_{x,y},c^{*}_{x,y}\right)+\frac{\lambda}{N_{pos}}\sum_{x,y}I_{\{c^{*}_{x,y}>0\}}L_{reg}\left(t_{x,y},t^{*}_{x,y}\right)$$
$$+\frac{1}{N_{pos}}\sum_{x,y}I_{\{c^{*}_{x,y}>0\}}L_{centerness}\left(s_{x,y},s^{*}_{x,y}\right) \tag{3-13}$$

式中，L_{cls} 为二元交叉熵损失和 Focal Loss；L_{reg} 为 IOU 损失；$L_{centerness}$ 为二元交叉熵损失；$p_{x,y}$ 为预测分类概率；$t_{x,y}$ 为预测目标边界框信息；$s_{x,y}$ 为预测的 Centerness；N_{pos} 为正样本的数量；I 为指示函数，当匹配正样本时为 1，否则为 0。

3.4 实验结果与分析

3.4.1 数据集以及实验条件介绍

目前尚未有公开的车辆部件检测数据集，故作者采集了大量复杂场景下的车辆部件图片建立了数据集。自建车辆部件检测数据集主要包括两部分：大部分与重庆云石高科技有限公司合作车检项目采集；另一小部分通过互联网爬虫获取，包括 VOC 数据集[74]和 COCO 数据集[75]的部分车辆类别图片。本人对自建数据集进行了严格的数据清洗工作，确保数据充足，场景丰富且具有普遍性。因为图像之间有明显差异，且应用场景也不近相同。所以本节将分为两个数据集来进行车辆部件检测测试。数据集的展示如图 3-11。

(a)　　(b)

图 3-11 数据集的部分图片展示。 (a) VLC 数据集；(b) VDC 数据集

Fig.3-11 Part of the image display of the dataset. (a) VLC dataset; (b) VDC dataset

从图 3-11 可以看到，VLC 数据集的部件均是在车辆内部，近距离视角拍摄的，因此部件在图像中均是大尺度物体目标。有安全带、铭牌、摩托车、灭火器、安

全锤、轮胎、车架号、身份证、驾驶证和行车记录仪共 10 个类别，共计 45210 张图片。VDC 数据集的部件均是在车辆外部，远距离视角拍摄的，因此在图像中这些部件是密集的小目标对象。有车脸、车尾、车牌、车标、车灯、转向灯、车桨、尾翼、轮胎、排气孔、行李架、三角架和后视镜共 13 个类别，共计 43883 张图片。

本章实验部分所用的软硬件配置如下：

（1）中央处理器：Intel Core i7-8700K

（2）内存（RAM）：64.0GB

（3）显卡：GeForce RTX 2080 Ti，12.0GB

（4）依赖库：python3.8，torch1.3.0 等

（5）系统类型：Ubuntu 18.04

在训练阶段，对 EFDet-SPP 在 VDC 数据集上的损失收敛过程进行了可视化，这样可以直观的观察到一些超参数的设定。如图 3-12 所示。图中依次为三分支预测网络的分类损失、Center-ness 损失和回归损失以及总损失。

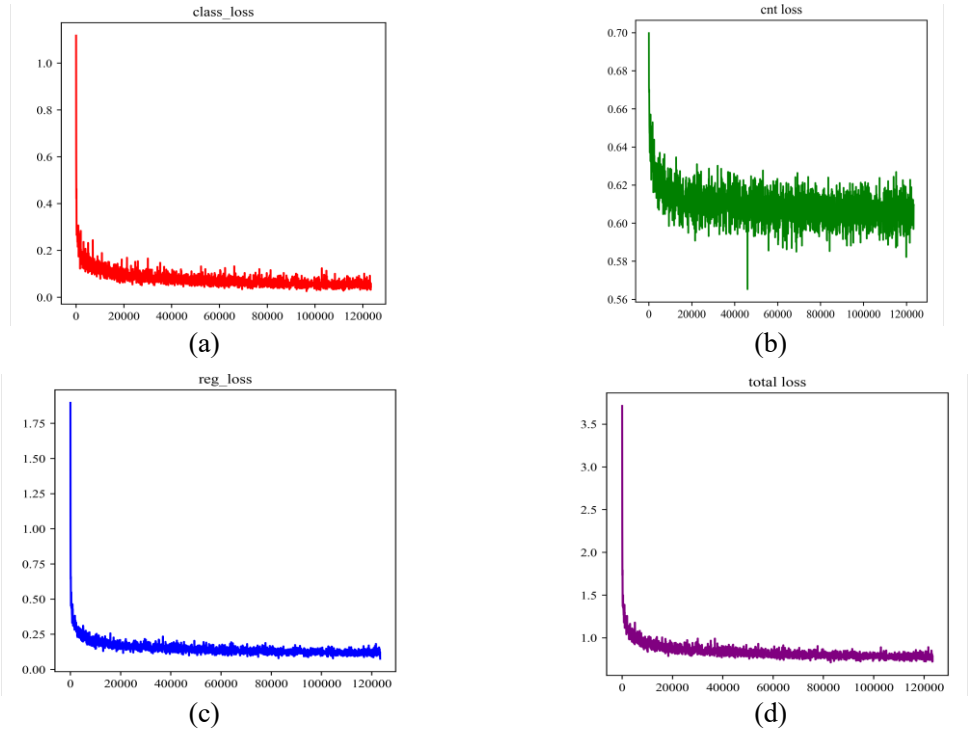


图 3-5 EFDet-SPP 损失收敛过程。(a) 分类损失；(b) Center-ness 损失；(c) 回归损失；(d) 总损失

Fig.3-5 The convergence process of EFDet-SPP loss. (a) Classification loss; (b) Center-ness loss; (c) Regression Loss; (d) Total loss

从图 3-12 可以看到，损失均在前 20000 次迭代中快速下降，在 20000-80000 区间逐渐收敛，整体趋势缓慢下降，在 80000 次迭代之后损失整体趋势平缓。因

此本章实验选择训练 35 个 epoch，bathsize 设置为 16，优化器为 Adamax，对于学习率使用阶梯型（step-based）衰减策略，设置初始学习率为 0.0001，如图 3-13 所示，随着不断迭代，学习率整体呈现阶梯状下降，这样能够在训练中动态调整学习率，以更好地适应模型的训练情况。

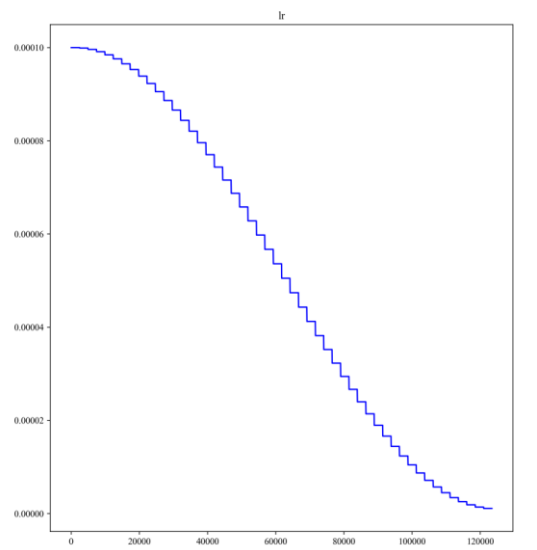


图 3-13 阶梯型衰减策略示意图

Fig.3-13 Illustration of step-based type attenuation strategy

本章的训练集和测试集是 VLC 和 VDC 数据集以 8：2 随机划分所得。整体训练耗费时间随着每次采用不同的优化模块而变化，整体上来看，一次训练耗时 25-30 个小时左右。

3.4.2 实验结果分析与展示

在本章内容的基础上，本节主要设计了两部分实验：本章提出的改进方式的消融实验和与其他主流算法的对比实验。

（1）消融实验

为了验证本章提出的各个部分进行优化的有效性，下面将通过图表的形式来展示实验结果。首先在 VLC 和 VDC 数据集上，本节对整个 EFDet-SPP 网络对各个部件的检测性能进行了测试。表 3-3 给出了每个类别相应的 AP 和 EFDet-SPP 的 mAP。由表 3-3 可以看出车牌、车脸、后视镜和急救锤这几个部件检测性能最好，排气孔、尾翼和转向灯这些部件检测性能有待提升。总体来说，EFDet-SPP 网络在两个数据集上的 mAP 均在 87%以上，在 VLC 数据集上甚至可以高达 94.8%，与 EfficientDet 网络相比有较大提升。

表 3-3 部件检测性能结果

Table3-3 Component detection performance results

VLC 数据集		VDC 数据集	
部件	AP（%）	部件	AP（%）
轮胎	70.09	车脸	97.75
铭牌	98.60	车牌	99.40
安全带	90.40	车标	90.67
灭火器	91.89	车灯	93.18
安全锤	98.98	车桨	90.04
车架号	95.74	尾翼	60.33
摩托车	93.47	轮胎	98.03
身份证	99.78	车尾	99.03
驾驶证	99.61	转向灯	85.96
行车记录仪	99.98	排气孔	81.76
/	/	行李架	84.39
/	/	三角架	99.16
/	/	后视镜	92.91
mAP	94.80	mAP	87.50

为了验证做出的优化到底能带来怎样的效果，本次以 EfficientDet 为基线模型，在检测难度较大的 VDC 数据集上，来进行消融实验。实验结果如表 3-4 所示，其中①表示基线模型，②表述数据增强方式，③表示空间金字塔池化模块，④表示纵向交叉跨层连接的 BiFPN，⑤表示无锚框预测方式。一般来说，模型参数量越大，网络越容易过拟合；FLOPs 越大，网络越难优化。

表 3-4 VDC 数据集上的消融实验

Table3-4 Ablation experiments on the VDC dataset

网络模型	模型参数量	FLOPs	mAP
基线模型	3.9M	**2.5B**	78.9
①+BiFPN 初始版本	4.0M	**2.5B**	78.0
①+纵向交叉跨层连接的 BiFPN	4.0M	2.6B	81.6
①+④+数据增强方式	4.0M	2.6B	82.1
①+②+④+空间金字塔池化模块	4.1M	2.7B	82.9
①+②+③+④+无锚框预测方式	**1.5M**	4.3B	**87.5**

表 3-4 内容为 VDC 数据集中每增加一个优化模块，其车辆部件检测的各项衡量指标的展示。可以看出，BiFPN 优化的初始方案是让基线模型的 mAP 降低了 0.9%，反而添加了纵向交叉连接的 BiFPN 模块的 mAP 提升了将近 3%，紧接着每增加一个优化模块，mAP 均有所提升。其中加入数据增强方式后，mAP 提升近 0.5%，之后加入空间金字塔池化模块，使得 mAP 提升 0.7%。此时平均精度仍处于较低状态，考虑是锚框的机制导致的网络模型在应用场景的不适问题，所以之后变为了无锚框预测方式，在平均精度方面有近 5%的提升同时模型参数量也大大减少。由此可知，本章所提出的每个优化在 VDC 数据集上都有一定的提升效果，其中纵向交叉跳层连接的 BiFPN 和无锚框预测方式对平均精度的提升效果最为明显。

对车辆部件检测而言，在高精度的同时兼顾和计算成本以保持高效往往是最需要的。为了进一步验证每个优化方式的性能，在 VLC 数据集，同一条件下，以模型参数量为 X 轴，mAP 为 Y 轴，对基线模型和其改进方案进行了实验并比较如图 3-14 所示，其中对于优化方式的标注与表 3-4 相同，在此不再赘述，可以看到效果最好的模型应该处于图像左上角。

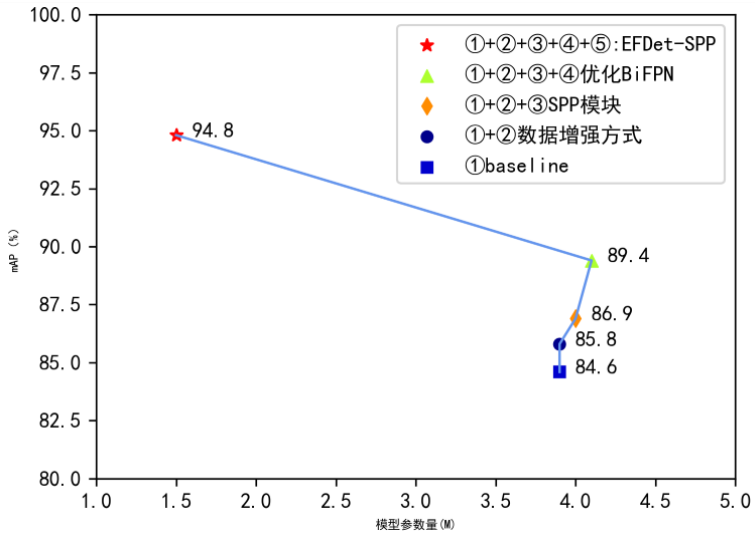


图 3-14 VLC 数据集上的消融实验

Fig.3-14 Ablation experiments on the VLC dataset

由图 3-14 可知， EFDet-SPP 在图像的左上角，性能效果是最好的，无锚框预测方式使其在平均精度和模型参数量方面均得到了较大的提升，同时对 BiFPN 的优化进一步提升了算法的性能。表 3-3、表 3-4 和图 3-14 的结果表明，本章提出的 EFDet-SPP 网络可以对车辆和车辆部件进行准确的检测判断，可以高效的开展安全防控工作等。

（2）对比实验

该部分实验旨在与其他优秀算法进行比较，观察各种算法之间的优缺点，验证 EFDet-SPP 的有效性。下面在 VLC 数据集上进行对比实验，实验结果如表 3-5 所示。

表 3-5 EFDet-SPP 与其他方法在 VLC 数据集上的对比

Table3-5 Comparison of EFDet-SPP with others on the VLC dataset

方法	Backbone	mAP（%）
Faster R-CNN[6]	VGG-16	87.3
SSD[14]	ResNet-101-SSD	79.8
YOLOv3[12]	DarkNet-53	80.3
RetinaNet[55]	ResNet-101-FPN	81.9
CornerNet[17]	Hourglass	87.6
EfficientDet[15]	EfficientNet	84.6
FCOS[22]	ResNet-101-FPN	90.2
Ours	EfficientNet	**94.8**

在表 3-5 中，Faster R-CNN 是两阶段目标检测算法，相比于单阶段目标检测算法，常常采用更大分辨率的图片作为输入。其达到了 87.3%mAP 的检测精度，相对来说是较高的。SSD、YOLOv3、RetinaNet 和 EfficientDet 均是单阶段目标检测方法，它们的检测精度较低，但是其只需要处理整个图像一次，算法的效率更高。CornerNet 和 FCOS 均是无锚框检测算法，在 mAP 方面表现较为突出，像 FCOS 达到了高达 90.2%的检测精度，分析原因是因为无锚框预测方式使得网络具有更强的场景适应性。虽然 EfficientDet 在 COCO 数据集上取得优异的效果，但是因为数据集标签框之间的巨大差异性，其并不能在车辆部件检测发挥较好的性能。反而本章提出的无锚框车辆部件检测网络 EFDet-SPP 取得了最好的效果，超过了基线模型 EfficientDet 的检测精度且有明显差距，相比于无锚框检测模型 FCOS 也提升近 5%。这样的增益恰恰验证了本章提出的优化方法的有效性。

综合以上 EFDet-SPP 的实验结果，证明了 EFDet-SPP 在车辆部件检测平均精度和模型参数量方面相比 EfficientDet 都有明显的提升。随着对基线模型一步步的优化，在车辆部件检测的性能指标上也有一定的增长，充分证明了数据增强方式、空间金字塔池化模块、纵向交叉跨层连接的 BiFPN 和无锚框预测方式，可以提升整体网络的车辆部件检测性能。

为了更加具体的展现本章提出的 EFDet-SPP 的车辆部件检测能力，将展示 EFDet-SPP 在难度较大的 VDC 数据集上的部分检测结果，如图 3-15 所示。

图 3-15 EFDet-SPP 在 VDC 数据集上的部分部件检测结果

Fig.3-15 Partial component detection results of EFDet-SPP on VDC dataset

3.5 本章小结

本章首先对 EfficientDet 算法的网络架构进行了简单的介绍，然后提出了一种基于 EfficientDet 的改进无锚框单阶段车辆部件检测网络 EFDet-SPP。EFDet-SPP 主要由特征提取网络、特征融合网络和分类定位网络组成，首先结合了 Mosaic 和复制粘贴的数据增强方法来增加小目标部件样本数量，然后为了有效地捕获高语义信息，在特征提取网络引入了空间金字塔池化，同时在特征融合网络 BiFPN 上增加纵向的交叉跨层连接数据流，平衡了自下而上路径的特征结点的输入数据流，

充分利用不同层级之间丰富的特征信息，从而提升了用于检测特征图的特征表达能力。在分类定位网络与 FCOS 目标检测模型相结合，改变了基于锚框的预测方式，转变为基于像素点预测，消除了与锚框相关的超参数从而减少了计算量，提高网络在不同场景下的适用性。为了验证该网络的有效性以及车辆部件数据集的稀缺，采集了各个地区的车大量车辆图片进行标注，建立了两个用于车辆部件检测的 VLC，VDC 数据集。实验表明 EFDet-SPP 实现了高效准确的车辆部件检测。

第 4 章 基于部件与全局特征的多粒度车辆重识别算法

第三章提出了车辆部件检测网络（EFDet-SPP），可以高效的识别出部件的类别与位置。而在车辆重识别任务中细节特征至关重要，由此以车辆部件检测网络 EFDet-SPP 为基础来提取部件的局部特征，在本章中提出了基于部件与全局特征的多粒度车辆重识别算法。

4.1 引言

车辆重识别在目标重识别领域备受关注，初期的重识别技术主要应用于行人重识别领域。随着国家对道路监控的愈加重视以及公安刑侦的迫切需要，车辆重识别逐渐引起关注。相较于行人重识别，车辆重识别问题的研究时间很短。车辆重识别的重难点：其一，由于不同拍摄视角的切换和光照等外在环境条件的改变，同一辆车在二维图像上的呈现会大不相同。其二，同一车型的车辆个体非常相似，不同款式或者车型的车辆个体在视觉上可能十分相似。车辆重识别与车牌检测、车型分类有所不同。由于车辆重识别不能仅依靠车牌这一唯一标识信息，因此需要更加精细的特征信息。全局特征很快就到达了精确率的瓶颈，因此需要使用更加精细的特征信息。提取局部特征与全局特征相结合可以帮助分辨相似车辆的细小差异，如对于车型和外观相同的车辆，其车辆外饰有特殊细节的情形。采用全局与局部特征相结合的方法是提高车辆重识别关键可分辨信息的重要途径，这种方法为冲破车辆重识别的瓶颈提供了很好的思路。

参考人类视觉中辨认一辆车取决于看到的极具辨识度的区域，本章着眼于寻找具有区分度的局部特征，并且不忽略整张车辆图像的全局信息。由此，提出了一种基于部件与全局特征的多粒度车辆重识别算法。该方法以 EFDet-SPP 网络来定位部件（车窗、车脸）以提取部件特征，其具体细节将在 4.2 节中具体介绍。在 4.3 节中，介绍了所做的相关实验并且分析了此方法的精度与效率。在 4.4 节中，对本章内容进行总结。

4.2 基于部件与全局特征的多粒度车辆重识别网络

在本节中，本文将介绍基于部件与全局特征的多粒度车辆重识别网络，网络的结构如图 4-1 所示。给定一张车辆的图像，由于待检测定位简化为车窗和车脸，且识别精确度较高，对第三章提出的车辆部件检测定位网络进行了简化，构建了车辆部件（车窗、车脸）检测定位模块；之后以 ResNet-50 的前四层为骨干网络模型，对全局支路、车窗局部支路和车脸局部支路提取相应的深度卷积特征从而令网络学习到一个由粗略到精细的结构化的特征表示。对于全局支路，采用 1×1、3×3、5×5、7×7 这 4 种不同尺度的卷积核来提取整张车辆图像的多尺度信息；

对于局部支路，采用多粒度划分策略，从垂直方向将特征图划分为多个条带，并对每个条带分别进行全局最大池化和 1×1 卷积。最后计算车辆图像与检索集中图像的余弦相似度，实现车辆重识别。

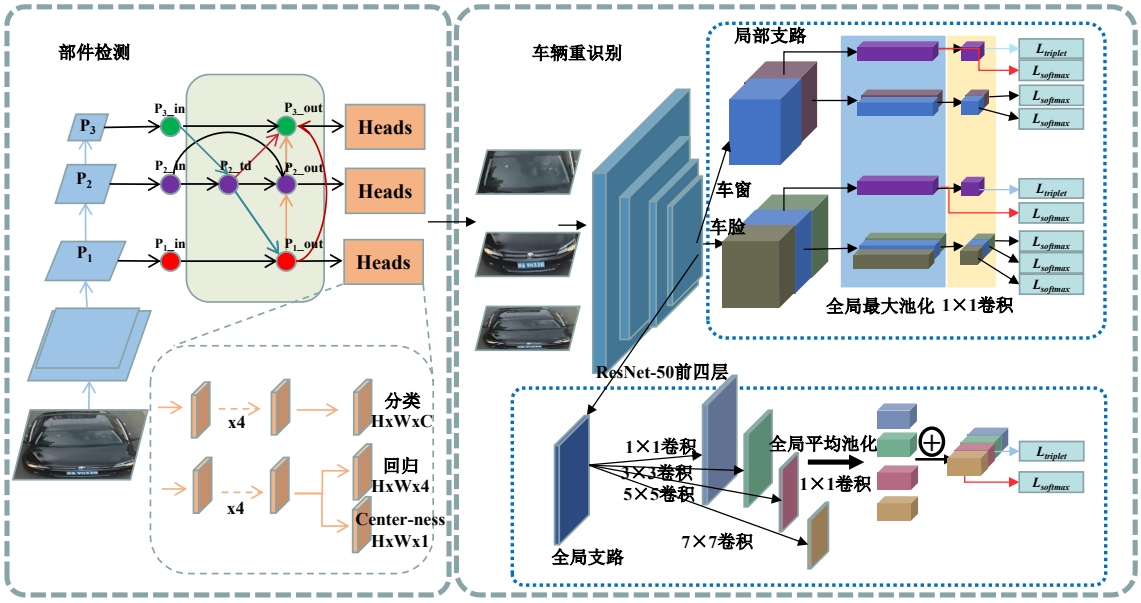


图 4-1 本章提出的网络结构，主要由三个分支组成：车窗、车脸局部支路和全局支路

Fig.4-1 The network structure proposed in this chapter is mainly composed of three branches： window, face local branch and global branch

4.2.1 部件检测定位网络

部件检测部分的作用是定位部件的局部图像以提取对应的局部特征。随着视角、光照和姿态的变化，像后视镜、尾灯和排气孔等小部件在查询集图像存在但是检索集图像却未能展显，对于局部特征的贡献会造成消极影响。由此本节仅选取了标志性较强且识别稳定的车窗和车脸部件作为局部特征，其图像如图 4-2 所示。

(a) (b)

图 4-2 局部部件示例。 (a) 车窗；(b) 车脸

Fig.4-2 Local component example. (a) Vehicle window; (b) Vehicle face

多分类且类间相似的检测任务需要复杂的模型来进行充分的特征提取融合来进行分类和定位。车辆部件检测任务中，部件的种类设计比较复杂且种类间极为相似，第三章中从精度和效率两个角度入手，构建了性能均衡的车辆部件检测网络结构 EFDet-SPP。从图 4-2 可以观察到车窗和车脸的特征属性区分度较大，该部件检测任务比较简单。综合衡量之下，对 EFDet-SPP 进行了精简，在保证精确度的同时尽可能的减少了一些需要大量计算的卷积操作。简化后的部件检测模块具体如图 4-3 所示。

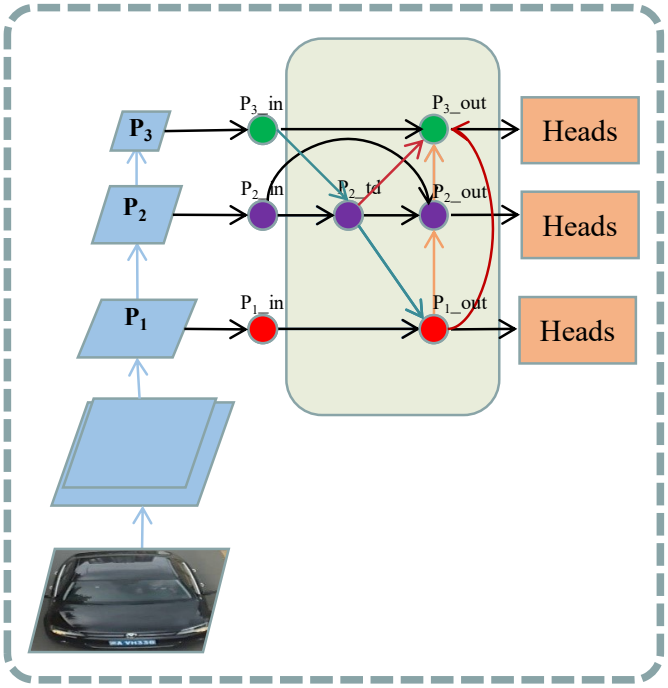


图 4-3 简化的 EFDet-SPP 网络结构

Fig.4-3 Simplified EFDet-SPP network structure

车辆重识别公共数据集的图像尺寸普遍偏小，因此网络的输入尺寸设定为 256×256。在骨干网络中，减去了最顶两层卷积层以及空间金字塔池化模块，同时只取 P_1、P_2 和 P_3 三个特征图来进行融合。对应的 BiFPN 网络也转变为了三层特征融合结构，保留了两条纵向的跨层连接（红色）。这里以 P_3 层为例，介绍此时特征融合的输出表达式如下：

$$P_3^{out}=Conv\left(\frac{\omega_1 P_3^{in}+\omega_2 Resize(P_2^{td})+\omega_3 Resize(P_2^{out})+\omega_4 Resize(Resize(P_1^{out}))}{\omega_1+\omega_2+\omega_3+\omega_4+\epsilon}\right) \tag{4-1}$$

式中，ω_1、ω_2、ω_3 和 ω_4 分别为 P_3^{out} 的输入权重。

4.2.2 全局支路

全局支路主要是针对整辆车进行特征提取，一般情况下多是采用 3×3 的卷积核。本节为了加强网络捕获全局特征的多尺度信息，其一，增加了 1×1 卷积核来促进通道间的交流和信息整合。此外，还在卷积层后添加了非线性激活函数，这可以在不改变特征图尺寸的情况下增强网络的非线性特性。其二，增加了 5×5、7×7 这两种不同尺度的卷积核来扩大感受野。共使用四种不同的卷积核来提取多尺度特征，有效提高了对全局多尺度信息的表征能力。为了有效的节约参数量，在使用 3×3、5×5 和 7×7 卷积核进行卷积之前使用 1×1 卷积进行降维处理，随后对降维后的特征图进行大尺度的卷积处理，网络结构如图 4-4 所示。

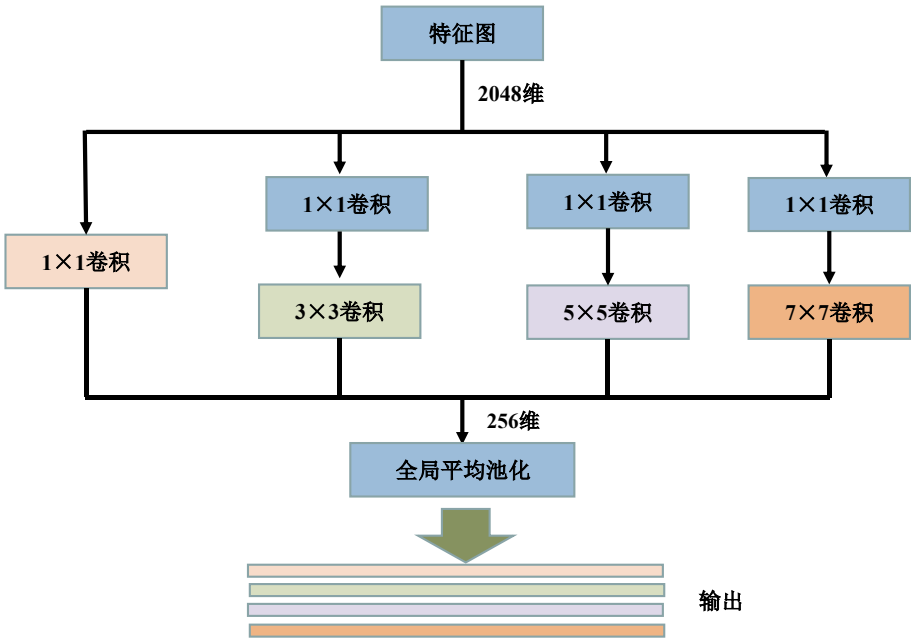


图 4-4 多尺度特征提取示意图

Fig.4-4 Illustration of multi-scale feature extraction

从图 4-4 可以看到，全局支路增加了 1×1、3×3、5×5、7×7 这 四种不同尺度的卷积核，分别输出尺度不同的特征图。当对输入特征图及进行 3×3、5×5、7×7 卷积之前，会先用 1×1 卷积进行降维处理使 2048 维特征降至 256 维，然后对于输出的四个特征图做全局平均池化（Global Average Pooling，GAP），最终将其进行叠加得到一个全局多尺度特征图。

4.2.3 局部支路

局部支路包括车窗局部支路和车脸局部支路，它们都没有使用下采样操作，以保留适合局部特征提取的感受野。为了捕捉更多的局部特征，将特征图沿垂直方向均匀地划分为多个条带，并对每个条带进行全局最大池化（Global Max

Pooling，GMP）和 1×1 卷积。多个条带表示不同的粒度，条带越多粒度越细。条带划分与原始图像的映射关系如图 4-5 所示。

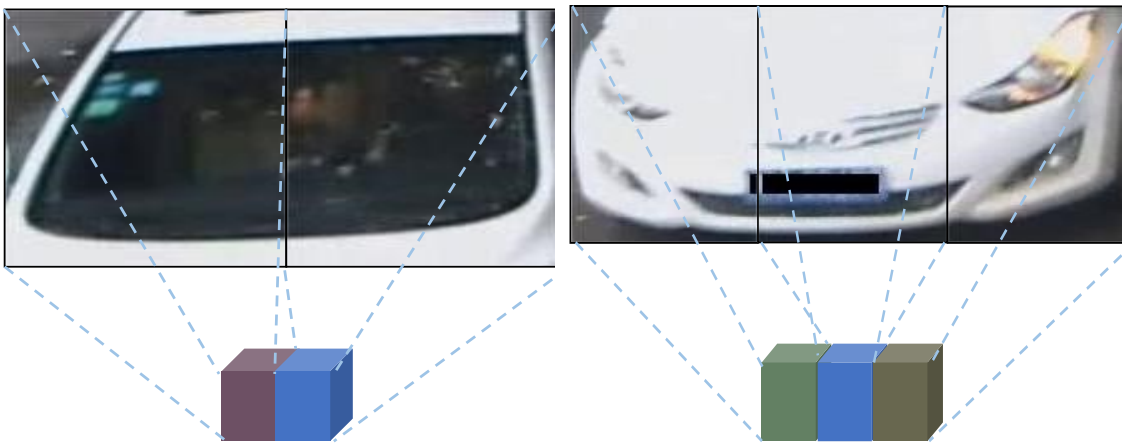


图 4-5 条带划分与原始图像的映射关系。 (a) 车窗局部支路；(b) 车脸局部支路

Fig.4-5 The mapping relationship between stripe division and the original image. (a) Partial branch of vehicle window; (b) Partial branch of vehicle face

对于车窗局部支路，将特征图划分为两个条带，更好的捕获细粒度细节，可以理解为车窗的左右两个部分。如图 4-5(a) 所示多数人习惯在车窗左右两部分悬挂吊坠等内饰物品，当车辆内饰有特殊细节时，对于车辆重识别的特征贡献度很大。

对于车脸局部支路，特征图被划分为三个条带，这样有利于网络来更好的捕获车灯、车牌和转向灯等辨识度较大的细粒度特征信息。如图 4-5(b) 所示可以理解为车的左车灯转向灯、右车灯转向灯和车牌区域三部分。

对于全局支路和局部支路，最终获得的特征信息，陈述于表 4-1，其中粒度表示分割条带数量。

表 4-1 三条支路的特征对比

Table4-1 Comparison of the characteristics for three branches

支路	粒度	特征图（H×W）	维度
全局支路	1	8×8	256
车窗局部支路	2	16×16	256*2+256
车脸局部支路	3	16×16	256*3+256

表 4-1 列出了三个支路的相关信息。在全局支路中，本节在共享的 ResNet-50 骨干网络之后采用步长为 2 的卷积层进行了一次降采样，在相应的输出特征上进行多尺度特征提取操作；对于两个局部支路，本节采用了全局最大池化层和 1×1 卷积层来处理输入特征图。这里的 1×1 卷积层包括归一化和 ReLU 激活函数，可以将 2048 维度特征降至 256 维，同时增强了网络的非线性特性。在测试阶段，本节将 3 个 256 维的全局特征向量和 5 个 256 维的局部特征向量叠加在一起，得到一个 2048 维的向量，作为车辆的特征表示。这个特征向量用于进行车辆的相似性搜索。

4.2.4 损失函数

为了释放该网络结构的学习表示的重识别能力，用来分类的 softmax 损失和度量学习的三元组损失将作为模型训练的损失函数。softmax 损失是 softmax 层和交叉熵损失组合而成的损失函数，三元组损失用来扩大类间距离，缩小类内距离，提升检索的准确性。在训练时，则是对同类损失做均值运算然后加权相加。

网络采用的 softmax 损失函数如下：

$$L_{softmax}=-\sum_{i}^{N} log\frac{e^{\omega_{y_i}^{T}f_i}}{\sum_{k=1}^{C}e^{\omega_k^{T}f_i}} \tag{4-2}$$

式中，W_k 对应着类别 k 的权值向量；N 表示训练过程中的批量大小；C 表示训练数据集中的类别总数。

与传统 softmax 损失不同的是，此处放弃了偏重以此来获得更佳的辨别性能。特征距离采用的余弦相似度，存在内积运算。若在内积计算后添加偏重，某些类别的特征可能会分布在原点附近，与其它类重叠[78]。这种情况会破坏某些类别的判别性。

网络采用的三元组损失如下：

$$L_{triplet}=-\sum_{i=1}^{P}\sum_{a=1}^{K}\left[\alpha+\max_{p=1\ldots K}\left\|f_a^{(i)}-f_p^{(i)}\right\|_2-\min_{\substack{n=1\ldots K\\ j=1\ldots P\\ j\neq i}}\left\|f_a^{(i)}-f_n^{(j)}\right\|_2\right]_+ \tag{4-3}$$

式中，$f_a^{(i)}$、$f_p^{(i)}$和 $f_n^{(i)}$为从真实框、正样本和负样本中提取的特征；α 是控制距离内外差异的超参数。这里的正样本和负样本指的是与真实框相同或不同的车辆。这里采用的是批量硬三元组损失[79]（Batch Hard Triplet Loss），是一个基于原始三元组损失的改进版本，这种改进的三重损失不仅增强了度量学习的鲁棒性，还进一步提高了模型的性能。

4.3 实验结果与分析

4.3.1 数据集介绍

数据集的质量和规模对于车辆重识别网络起着至关重要的作用，现在的车辆重识别数据集图片大多来源于现实场景下的交通监控系统，其需要数据量庞大且数据极难获取，所以它的数据集比较少。目前使用频率最高的数据集有 VeRi-776 数据集[80]和 VehicleID 数据集[81]，这两个数据集具有较高的质量和规模，能够支持对车辆重识别算法进行有效的评估和优化。因此，本节的实验主要在这两个数据集上展开。

（1） VeRi-776 数据集

VeRi-776 数据集是由 Liu X 等人[80]在 2016 年提出的现实世界包含多个车辆数据的一个大规模数据集。它的数据来源于 20 台摄像机在 1 平方公里的城市区域内 24 小时拍摄采集的图片。数据集图像共包含 776 辆车的超过 50000 张图像，其中包含 576 辆车的 37778 张图像作为训练集，用于训练车辆重识别网络以达到收敛，包含 200 辆车的 11579 张图像作为测试集，用于检验网络的各种性能指标，剩余的 1678 张图像作为查询集。这些车辆图像都是从 2 到 18 个角度拍摄得到，具有不同的光照条件以及遮挡复杂的背景信息。数据集中的图像标注了足够的属性和时空信息，例如车牌信息，拍摄的时间戳以及相邻相机之间的距离。尽管车辆的属性和时空信息对于车辆重识别算法具有很大的帮助，但在现实应用中，它们并不常用于训练。这是因为这些信息需要人工标注，这不仅增加了数据集的制作成本，还可能降低车辆重识别网络的泛化能力和实用性。因此，在本文的车辆重识别算法中，没有使用这些属性和时空信息。VeRi-776 数据集中训练集与测试集的部分图像如图 4-6 所示，其中前两行和第三行分别展示了训练集和测试集的部分图像。

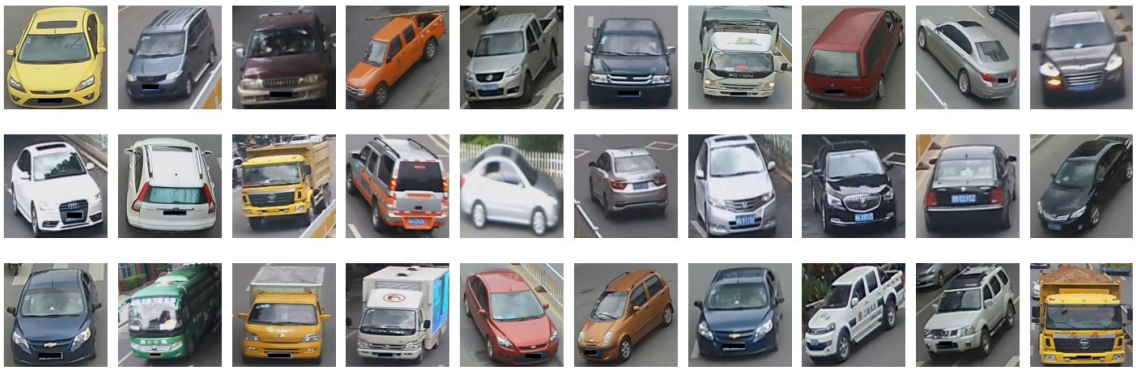


图 4-6 VeRi-776 数据集部分图像展示

Fig.4-6 Partial image display of VeRi-776 dataset

（2）VehicleID 数据集

VehicleID 数据集[81]也是车辆重识别领域的一个大规模数据集。相较于 VeRi-776 数据集，VehicleID 数据集包含了前后两种视角下 26267 辆车的 221763 张图片，并且标注了车辆的车型信息，其中训练集包含 13134 辆车的 110178 张图片；测试集包含了 13133 辆车 111585 张图片，并且有着小中大三个验证集，能满足模型对于测试集不同大小的需求。如图 4-7 所示，是对 VehicleID 数据集中部分数据样本的展示。

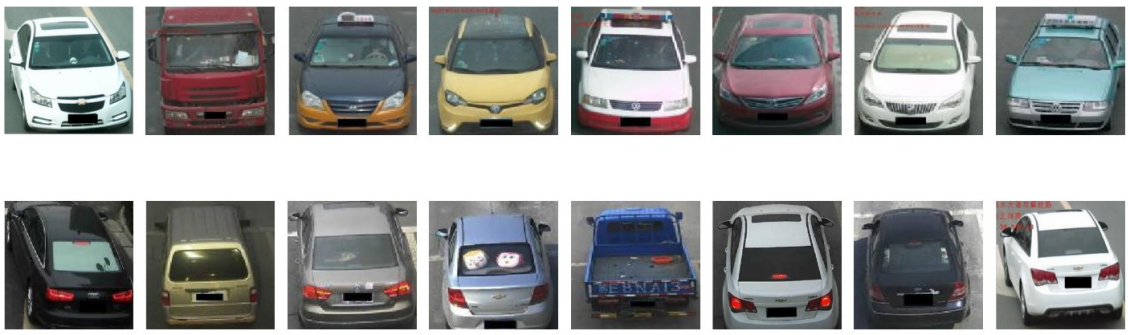


图 4-7 VehicleID 数据集部分图像展示

Fig.4-7 Partial image display of VehicleID dataset

表 4-2 是 VeRi-776 数据集和 VehicleID 数据集在样本数量、车辆数量、视角个数和时空信息四个方面的详细对比结果。从表 4-2 可以发现，VehicleID 数据集有更多的数据样本，有利于网络训练。但 VeRi-776 数据集有更多的视角变化，更具真实性且规模可观。

表 4-2 数据集比较

Table4-2 Comparison of dataset

数据集	样本数量	车辆数量	视角个数	时空信息
VeRi-776	50000	776	2-18	是
VehicleID	221763	26267	2	否

4.3.2 实验条件与训练过程

实验所需的硬件条件大致与 3.4.1 节中相同，此处就不再进行赘述。

首先是进行车辆重识别网络训练前的一些数据处理。因为本章的目的是希望通过改进后的车辆部件检测网络提取局部部件图像，从而和全局特征结合改进车辆重识别的性能。所以在车辆数据集上验证了简化后的 EFDet-SPP 车辆部件检测器的性能，图 4-8 显示了车窗、车脸部件检测结果。

图 4-8 车辆部件检测模块检测结果

Fig.4-8 Vehicle component detection module test results

从图 4-8 可见，通过训练好的简化 EFDet-SPP 检测器，可以很好的检测出车窗、车脸等局部感兴趣区域。将车窗、车脸和整车图像进行叠加同时将训练图像尺寸调整至 256×256 大小就获得了重识别网络的输入。

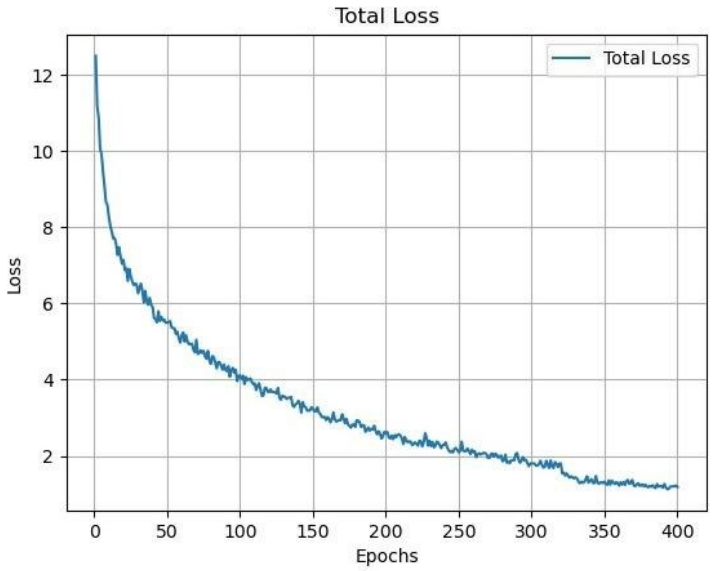


图 4-9 车辆重识别网络的损失收敛过程

Fig.4-9 The loss convergence process of vehicle re-identification network

如图 4-9 所示，其展示了重识别网络在 VeRi-776 数据集上的训练损失收敛过程。可以观察到，总损失在刚开始训练的时候就迅速下降，而后在大约 350 个 epoch 之后下降速度减缓；之后的 50 个 epoch 中，损失逐渐收敛，并最终趋于稳定。因此本章重识别网络训练时设置 400 个 epoch 结束，使用 Adam 优化器训练模

型，并将初始学习率 lr 设为 0.0002，动量为 0.9，使用随机擦除进行数据增强，整体训练时间在 30-40 个小时左右。

4.3.3 实验结果展示与分析

下面基于本章的内容，本节共设计两个实验，包括：

（1）第一个实验主要是整体框架的性能测试以及部件检测部分与第三章中的算法的对比，主要为了验证本章的重识别网络是否可以有效的完成任务。本章将此部分称为框架实验；

（2）第二个实验是与其他优秀算法进行对比实验，通过与车辆重识别任务中主流算法对比，验证了该算法的优异性，同时针对本章做出局部多粒度支路和全局支路对比证明局部特征与全局特征结合的方式带来的提升幅度。本章将该部分简称为对比试验。

下面对每个实验的结果进行具体分析。

（1）框架实验

对于该部分实验，主要是为了展示本章提出的车辆重识别网络及其部件检测模块的性能。本章在 VeRi-776 车辆重识别数据集上进行了大量实验，如图 4-10 所示，为车辆重识别的 mAP、Rank-1、Rank-3、Rank-5 和 Rank-10 随着不断训练的增长曲线。

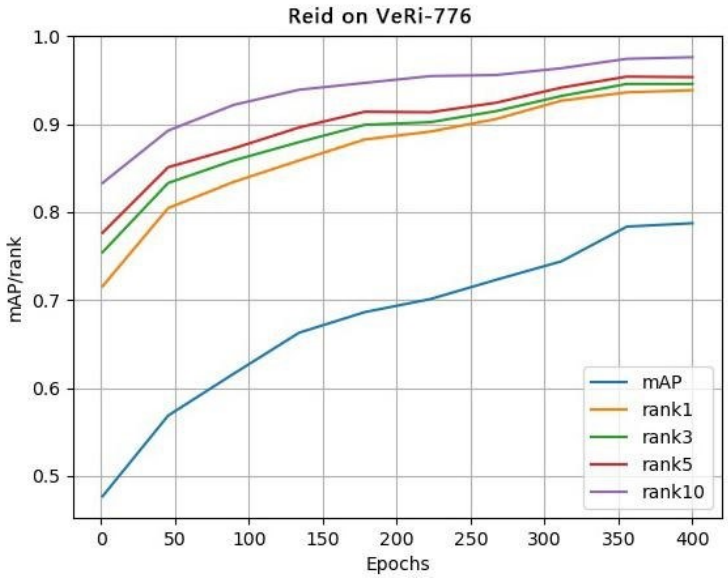


图 4-10 mAP、Rank 指标增长曲线

Fig.4-10 Growth curve of mAP and Rank

从图 4-10 可以看出，本章车辆重识别网络的 mAP 在 VeRi-776 数据集上接近 0.8，达到了极高的水平，并且其 Rank-1、Rank-3、Rank-5 和 Rank-10 等指标几乎

高达 0.95 以上。然后对于 VDC 和 VeRi-776 数据集进行处理得到车窗和车脸部件检测数据源，下面将对本章的部件检测部分和第三章中的车辆部件检测算法 EFDet-SPP 进行了比较，实验结果如表 4-3 所示：

表 4-3 EFDet-SPP 简化的有效性

Table4-3 The effectiveness of EFDet-SPP simplification

检测网络	模型参数量	FLOPs	mAP
EFDet-SPP	1.5M	4.3B	98.7
简化的 EFDet-SPP	1.2M	2.9B	97.9

如表 4-3 所示，简化后的 EFDet-SPP 与 EFDet-SPP 的检测性能（mAP）相差无几，但是 EFDet-SPP 的模型参数量与 FLOPs 要高出很多，证明了对 EFDet-SPP 简化的有效性。综上所述，本章提出的车辆重识别算法采用部件的多粒度局部特征和全局特征相结合的方式，可以在车辆重识别任务中实现高精度识别，同时对 EFDet-SPP 的简化也发挥了其减小模型占用内存和计算量的作用。

（2）对比实验

该部分实验的主要目的是将本章提出的车辆重识别算法与其他主流算法进行对比，并进行支路对比实验。本章首先在 VeRi-776 数据集上进行了大量实验，对于 VeRi-776 数据集，采用 mAP、Rank-1、Rank-5 指标评测。结果如表 4-4 所示。

表 4-4 VeRi-776 数据集重识别结果对比

Table4-4 Comparison of re-identification results of VeRi-776 dataset

Methods	Rank-1	Rank-5	mAP
LOMO[82]	23.8	46.4	9.8
PROVID[83]	61.4	78.8	27.8
CCL[81]	64.1	82.7	38.4
ICV[84]	96.2	99.0	59.5
AAVER[85]	89.0	94.7	61.2
MTCRO[86]	87.9	94.3	62.6
MGL[87]	86.1	96.2	65.0
VANET[88]	89.7	95.9	66.3
PCRNet[89]	95.4	98.4	78.6
MCRL[90]	96.1	99.4	81.1
OURS	93.9	95.4	78.9

由表4-4可得，在VeRi-776数据集上。最终网络训练结果mAP为78.9，Rank-1为93.9，Rank-5为95.4；于以前的LOMO、CCL和ICV等方法相比均有跨越性的提升；相较于VANET方法mAP提高12.6%，Rank-1提高4.2%，Rank-5相差无几；相较于PCRNet方法mAP提高0.3%；与目前在该数据集上表现优秀的MCRL算法mAP无较大差距；因此本章提出的车辆重识别算法与大部分优秀方法相比有一定的提高，从整体上来说该算法的车辆重识别能力是相对有竞争力的。

为了测试本章算法的鲁棒性，下面在VehicleID数据集的三个测试子集（小、中、大）上进行了对比实验。在重识别任务中，Rank-1更为重要，要求越大越好。Rank-1是指在检索结果中，第一个返回的车辆图像是正确的概率。因此评测指标主要采用Rank-1、Rank-5和mAP，实验结果如表4-5所示：

表4-5 VehicleID数据集重识别结果对比

Table4-5 Comparison of re-identification results of VehicleID dataset

	Test Size=800			Test Size=1600			Test Size=2400		
Methods	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP
LOMO[82]	19.8	32.0	/	18.9	29.2	/	15.3	25.3	/
PROVID[83]	49.5	68.1	/	44.6	64.6	/	40.0	60.4	/
GoogLeNet[77]	47.9	67.4	46.2	43.4	63.9	44.1	38.3	59.4	38.1
CCL[81]	53.9	69.3	/	51.2	67.0	/	46.9	62.9	/
AAVER[85]	74.7	93.8	/	68.6	90.0	/	63.5	85.6	/
MTCRO[86]	89.0	94.8	/	90.4	94.7	/	85.1	93.1	/
MGL[87]	79.6	94.0	/	76.2	91.2	/	73.0	88.2	/
VANET[88]	83.3	96.0	/	81.1	94.7	/	77.2	92.9	/
PCRNet[89]	86.6	98.1	/	82.2	96.3	/	80.4	94.2	/
MCRL[90]	88.1	97.4	92.0	83.3	95.8	88.4	80.0	80.7	86.0
OURS	89.6	96.3	90.7	91.3	93.3	87.6	85.9	89.2	82.6

通过表4-5可知，在VehicleID数据集的三种测试集上与其他主流算法相比，本章提出的算法在Rank-1指标均表现更优。这说明列表第一位是正确匹配的概率更大，与多粒度划分的局部特征息息相关。同时实验结果相对于大部分主流算办法均有大幅度提升，如在大测试集上相较于GoogLeNet方法mAP提高近50%。在三种尺度测试集，与表4-5中最优算法相比，mAP低了1%-4%，但是整体上Rank-1和Rank-5具有更好的效果，整体上来说有一定的准确率优势。

下面是对本章车辆重识别的局部支路和全局支路进行对比实验，表4-6为在VeRi数据集上的mAP和Rank-1对比表。

表 4-6 支路融合实验

Table4-6 Branch fusion experiment

算法	Rank-1	Rank-5	mAP
全局支路	87.7	89.6	71.2
局部支路	44.3	47.6	42.9
全局+局部支路	93.9	95.4	78.9

由表 4-6 可知，因为拍摄角度不同，单独使用局部支路难以发挥出多粒度局部特征的贡献。但是结合车辆部件的局部特征和全局特征，比仅使用全局特征更加具有优势。在 VeRi-776 数据集上，结合全局支路和局部支路的 mAP、Rank-1 和 Rank-5 比仅使用全局支路提升了 7.7%、6.2%和 5.8%，达到一个更好的准确率。

综上所述，本章的车辆重识别算法不仅能学习整辆车的全局信息，还可以学习多粒度划分的部件局部信息，结合两者的优势，在 VeRi-776 数据集和 VehicleID 数据集上都取得了相对具有竞争力的效果。最后对于本章算法的重识别进行了可视化如图 4-11 所示，其中排序为绿色为匹配成功，红色为匹配错误。

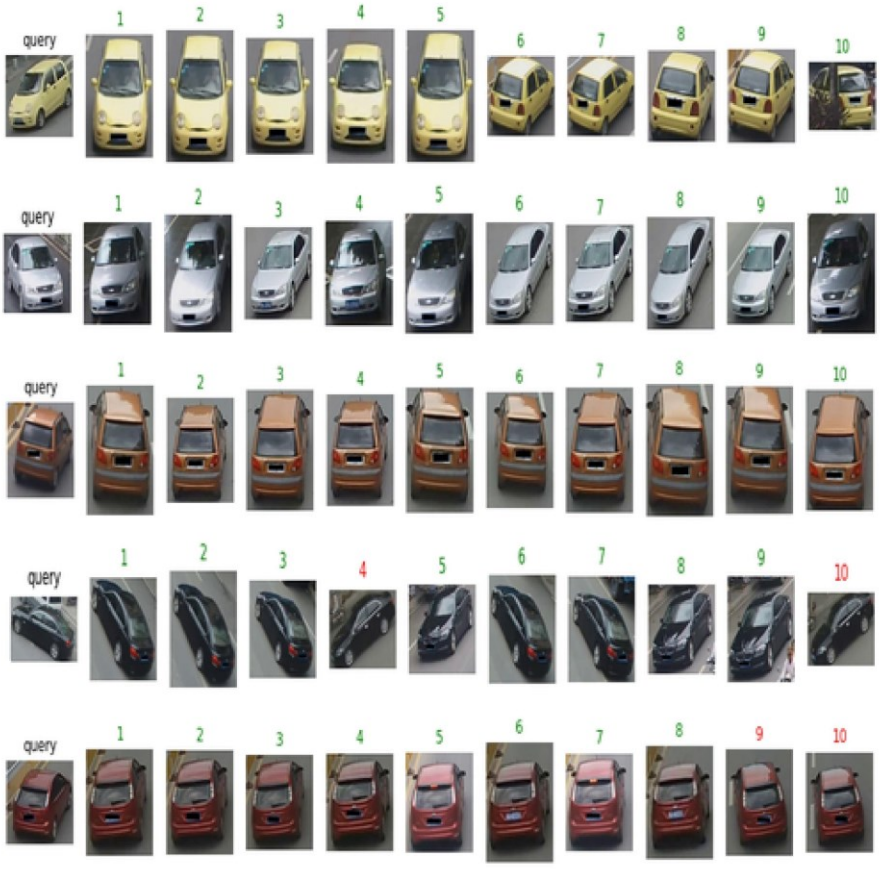


图 4-11 VeRi-776 数据集重识别结果图

Fig.4-11 Re-identification results of VeRi-776 dataset

4.4 本章小结

本章主要针对车辆重识别任务中不同车辆外观几乎相同的问题，提出一种结合全局特征与多粒度划分的部件局部特征的车辆重识别网络。在本章中，展示了重识别网络的整体流程，包括部件检测定位网络、全局支路、局部支路和损失函数。其中部件检测定位网络为优化的 EFDet-SPP；全局支路采用了四个不同大小卷积核来进行多尺度提取全局特征；局部支路设计了一种特征学习策略，将判别信息划分为各种粒度特征。同时不在语义区域进行学习，而是从垂直方向将图像特征均匀地划分为若干条纹，改变不同局部分支中的条纹数量，从而获得多粒度的局部特征表示；结合 softmax 损失函数（去偏重）和批量硬三元组损失函数进行优化。最后，根据各个车辆获得的特征图之间的余弦相似度进行排列从而完成了本章车辆重识别网络的算法设计，同时在公开数据集上做了大量对比实验，证明了该方法的有效性。

第 5 章 总结与展望

5.1 主要结论

随着计算机视觉技术和深度学习的蓬勃发展，基于深度学习的车辆部件检测与重识别成为当前研究的焦点。这一技术领域是计算机视觉技术的重要分支之一，其广泛应用于智能交通系统、智能停车、安防监控和自动驾驶等多个领域，正展现出无限的前景和实际价值。本文对深度学习网络 EfficientDet 进行优化，提出了一种高效准确的车辆部件检测网络 EFDet-SPP，并用此网络提取局部图像，利用部件的多粒度局部特征与全局特征实现了较为准确的车辆重识别。下面从车辆部件检测和车辆重识别两个方面对本文所做的主要工作与贡献进行概述：

首先在车辆部件检测中：

（1）在对特征融合的研究中，受到 BiFPN 跨层连接以及残差网络的思想启发，设计了一种纵向交叉跨层连接的 BiFPN，综合平衡了高层结点和底层结点的输入数据流，缩短了底层特征图到顶层特征图的距离，更好的融合了底层和顶层特征信息，实现了更好的信息交互。

（2）本文通过对有锚框和无锚框两种常用预测方式的优劣性分析，结合车辆部件检测任务的需求和特点，将基于锚框预测转变为基于像素点预测，消除了与锚框相关的超参数从而减少了计算量，一定程度上提高了在车辆部件检测场景下的适用性。

（3）本文考虑到车辆部件的尺寸特点，在数据输入时，结合了 Mosaic 和复制粘贴两种数据增强方式来平衡样本，增强网络泛化能力。同时在骨干网络引入了空间金字塔池化来有效的捕获高语义信息。

（4）由于数据集的缺少，本文建立了两个车辆部件数据集 VLC 和 VDC，确保数据充足，场景丰富。并且通过实验验证了本文设计优化方法的合理性和有效性。此外，与其他经典目标检测模型 YOLOv3 和 RetinaNet 在相同的数据集上进行实验比较，本文提出的检测模型也展现出了较好的性能表现。

其次在车辆重识别中：

（1）本文着眼于寻找具有区分度的局部特征，并且不忽略整张车辆图像的全局信息，提出了一种基于部件与全局特征的多粒度车辆重识别算法。使用简化的 EFDet-SPP 捕获部件图像来提取局部特征，同时对整幅图像进行多尺度特征提取全局特征。因此，这不仅学习到了车辆的整体特征，也对区分度大的部件局部特征进行了有效的学习，并通过实验对比验证了本文车辆重识别算法的合理性和有效性。

（2）在对局部特征进行提取时，本文着眼于更加细粒度的特征，设计了一种新颖的特征学习策略。即将判别信息划分为各种粒度特征，同时不在语义区域进行学习，而是从垂直方向将图像特征均匀地划分为若干条纹，改变不同局部分支中的条纹数，以获得多粒度的局部特征表示。通过实验验证了这种特征学习策略的有效性。

5.2 研究展望

本文对车辆部件检测与重识别做了算法改进，并在部分数据集上取得了较好的效果，但是仍然能发现不少问题，未来需要持续优化和改善：

（1）数据集的扩充和多样化：本文提出的 EFDet-SPP 车辆部件检测网络是基于 VLC 和 VDC 进行实验验证的，目前，车辆部件检测与重识别的数据集相比其他计算机视觉任务是较少并规模不大，并缺乏多样性和代表性，难以覆盖不同场景和应用的需求。因此，未来需要构建更大规模，更丰富，更具挑战性的数据集，以提高车辆部件检测与重识别的泛化能力和实用性。

（2）模型的轻量化和高效化：车辆部件检测与重识别的模型通常需要处理大量的高维特征，导致计算量和存储空间的增加，限制了模型的部署和应用。因此，未来需要设计更轻量化，更高效化的模型，以降低计算成本和资源消耗，提高模型的速度和性能。

（3）多模态和多任务的融合：本文在车辆重识别中只在图像数据上考虑了局部特征和全局特征。然而车辆重识别的任务不仅仅是识别同一辆车辆，还需要考虑车辆的属性，如颜色、型号、品牌和车牌等；车辆的行为，如速度、方向和轨迹等。因此，未来可以利用多模态的信息，如图像、视频、文本和声音等，以及多任务的学习，如检测、分类和跟踪等，以提高车辆重识别的精度和鲁棒性。

参考文献

[1] 马红丽. 智能网联汽车全力智慧化[J]. 中国信息界, 2021(06): 51-54.

[2] 许洁琼. 基于视频图像处理的车辆检测与跟踪方法研究[D]. 青岛: 中国海洋大学, 2012.

[3] 王圣男, 郁梅, 蒋刚毅. 智能交通系统中基于视频图像处理的车辆检测与跟踪方法综述[J]. 计算机应用研究, 2005(09): 9-14.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.

[5] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2015: 1440-1448.

[6] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[7] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 6154-6162.

[8] Li Y, Chen Y, Wang N, et al. Scale-aware trident networks for object detection[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 6054-6063.

[9] Qin Z, Li Z, Zhang Z, et al. ThunderNet: Towards real-time generic object detection on mobile devices[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 6718-6727.

[10] Lin T-Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 2117-2125.

[11] Bochkovskiy A, Wang C-Y, Liao H-Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[12] Farhadi A, Redmon J. Yolov3: An incremental improvement[C]. Computer vision and pattern recognition, 2018: 1-6.

[13] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 7263-7271.

[14] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016: 21-37.

[15] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 10781-10790.

[16] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.

[17] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]. Proceedings of the European conference on computer vision (ECCV), 2018: 734-750.

[18] Law H, Teng Y, Russakovsky O, et al. Cornernet-lite: Efficient keypoint based object detection[J]. arXiv preprint arXiv:1904.08900, 2019.

[19] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 6569-6578.

[20] Zhou X, Zhuo J, Krahenbuhl P. Bottom-up object detection by grouping extreme and center points[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 850-859.

[21] Yang Z, Liu S, Hu H, et al. Reppoints: Point set representation for object detection[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 9657-9666.

[22] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 9627-9636.

[23] 高磊. 基于光流的动态场景中运动车辆检测与跟踪算法研究[D]. 合肥: 中国科学技术大学, 2014.

[24] Hilario C, Collado J, Armingol J M, et al. Pyramidal image analysis for vehicle detection[C]. IEEE Proceedings. Intelligent Vehicles Symposium, 2005., 2005: 88-93.

[25] 宋晓琳, 邬紫阳, 张伟伟. 基于阴影和类 Haar 特征的动态车辆检测[J]. 电子测量与仪器学报, 2015, 29(09): 1340-1347.

[26] Cao X, Wu C, Yan P, et al. Linear SVM classification using boosting HOG features for vehicle detection in low-altitude airborne videos[C]. 2011 18th IEEE international conference on image processing, 2011: 2421-2424.

[27] Guo E, Bai L, Zhang Y, et al. Vehicle detection based on superpixel and improved hog in aerial images[C]. Image and Graphics: 9th International Conference, ICIG 2017, Shanghai, China, September 13-15, 2017, Revised Selected Papers, Part I 9, 2017: 362-373.

[28] Laopracha N, Sunat K. Comparative study of computational time that HOG-based features used for vehicle detection[C]. Recent Advances in Information and Communication Technology 2017: Proceedings of the 13th International Conference on Computing and Information Technology (IC2IT), 2018: 275-284.

[29] Pan C, Sun M, Yan Z. The study on vehicle detection based on dpm in traffic scenes[C]. Frontier Computing: Theory, Technologies and Applications FC 2016 5, 2018: 19-27.

[30] Chávez-Aragón A, Laganiere R, Payeur P. Vision-based detection and labelling of multiple vehicle parts[C]. 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2011: 1273-1278.

[31] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]. Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, 2001: I-I.

[32] Lowe D G. Object recognition from local scale-invariant features[C]. Proceedings of the seventh IEEE international conference on computer vision, 1999: 1150-1157.

[33] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), 2005: 886-893.

[34] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(7): 971-987.

[35] Prakash S. False mapped feature removal in spin images based 3D ear recognition[C]. 2016 3rd international conference on signal processing and integrated networks (SPIN), 2016: 620-623.

[36] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[J]. Lecture notes in computer science, 2006, 3951: 404-417.

[37] Laptev I, Lindeberg T. On space-time interest points[J]. International journal of computer vision, 2005, 64(2-3): 107-124.

[38] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103: 60-79.

[39] Wang H, Wang Y, Zhou Z, et al. Cosface: Large margin cosine loss for deep face recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 5265-5274.

[40] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of machine learning research, 2009, 10(2).

[41] Zhang Y, Liu D, Zha Z-J. Improving triplet-wise training of convolutional neural network for vehicle re-identification[C]. 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017: 1386-1391.

[42] Sun Y, Cheng C, Zhang Y, et al. Circle loss: A unified perspective of pair similarity optimization[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 6398-6407.

[43] Ge W. Deep metric learning with hierarchical triplet loss[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 269-285.

[44] Zhong Z, Zheng L, Cao D, et al. Re-ranking person re-identification with k-reciprocal encoding[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1318-1327.

[45] Lecun Y. Generalization and network design strategies[J]. Connectionism in perspective, 1989, 19(143-155): 18.

[46] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 1-9.

[47] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[48] 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, 2016: 770-778.

[49] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 4700-4708.

[50] Meng D, Li L, Liu X, et al. Parsing-based view-aware embedding network for vehicle re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 7103-7112.

[51] Xu B, Wang N, Chen T, et al. Empirical evaluation of rectified activations in convolutional network[J]. arXiv preprint arXiv:1505.00853, 2015.

[52] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[53] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[54] 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.

[55] Lin T-Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE international conference on computer vision, 2017: 2980-2988.

[56] Liu S, Huang D. Receptive field block net for accurate and fast object detection[C]. Proceedings of the European conference on computer vision (ECCV), 2018: 385-400.

[57] Devries T, Taylor G W. Improved regularization of convolutional neural networks with cutout[J]. arXiv preprint arXiv:1708.04552, 2017.

[58] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[C]. Proceedings of the AAAI conference on artificial intelligence, 2020: 13001-13008.

[59] Kisantal M, Wojna Z, Murawski J, et al. Augmentation for small object detection[J]. arXiv preprint arXiv:1902.07296, 2019.

[60] Neubeck A, Van Gool L. Efficient non-maximum suppression[C]. 18th international conference on pattern recognition (ICPR'06), 2006: 850-855.

[61] Gidaris S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware cnn model[C]. Proceedings of the IEEE international conference on computer vision, 2015: 1134-1142.

[62] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 8759-8768.

[63] Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv:1911.09516, 2019.

[64] Zheng W-S, Gong S, Xiang T. Reidentification by relative distance comparison[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(3): 653-668.

[65] Koestinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints[C]. 2012 IEEE conference on computer vision and pattern recognition, 2012: 2288-2295.

[66] Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 403-412.

[67] Lu C. Shannon equations reform and applications[J]. BUSEFAL, 1990, 44: 45-52.

[68] Yu J, Jiang Y, Wang Z, et al. Unitbox: An advanced object detection network[C]. Proceedings of the 24th ACM international conference on Multimedia, 2016: 516-520.

[69] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 658-666.

[70] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]. Proceedings of the AAAI conference on artificial intelligence, 2020: 12993-13000.

[71] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 815-823.

[72] Zoph B, Cubuk E D, Ghiasi G, et al. Learning data augmentation strategies for object detection[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, 2020: 566-583.

[73] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]. International conference on machine learning, 2019: 6105-6114.

[74] Everingham M, Van Gool L, Williams C K, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2009, 88: 303-308.

[75] Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, 2014: 740-755.

[76] Yu F, Chen H, Wang X, et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 2636-2645.

[77] Yang L, Luo P, Change Loy C, et al. A large-scale car dataset for fine-grained categorization and verification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 3973-3981.

[78] Wang F, Xiang X, Cheng J, et al. Normface: L2 hypersphere embedding for face verification[C]. Proceedings of the 25th ACM international conference on Multimedia, 2017: 1041-1049.

[79] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

[80] Liu X, Liu W, Mei T, et al. A deep learning-based approach to progressive vehicle re-identification for urban surveillance[C]. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, 2016: 869-884.

[81] Liu H, Tian Y, Yang Y, et al. Deep relative distance learning: Tell the difference between similar vehicles[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 2167-2175.

[82] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 2197-2206.

[83] Liu X, Liu W, Mei T, et al. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance[J]. IEEE Transactions on Multimedia, 2017, 20(3): 645-658.

[84] Bai Y, Lou Y, Gao F, et al. Group-sensitive triplet embedding for vehicle reidentification[J]. IEEE Transactions on Multimedia, 2018, 20(9): 2385-2399.

[85] Khorramshahi P, Kumar A, Peri N, et al. A dual-path model with adaptive attention for vehicle re-identification[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 6132-6141.

[86] Xu D, Lang C, Feng S, et al. A framework with a multi-task CNN model joint with a re-ranking method for vehicle re-identification[C]. Proceedings of the 10th International Conference on Internet Multimedia Computing and Service, 2018: 1-7.

[87] Yang X, Lang C, Peng P, et al. Vehicle re-identification by multi-grain Learni[C]. 2019 IEEE International Conference on Image Processing (ICIP), 2019: 3113-3117.

[88] Chu R, Sun Y, Li Y, et al. Vehicle re-identification with viewpoint-aware metric learning[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 8282-8291.

[89] Liu X, Liu W, Zheng J, et al. Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification[C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 907-915.

[90] Jin Y, Li C, Li Y, et al. Model latent views with multi-center metric learning for vehicle re-identification[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(3): 1919-1931.