



JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
06
VOL.27

ISSN1006-8961
CN11-3758/TB

2021

图像图形学 发展 年度报告



第27卷第6期（总第314期）
2022年6月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主 编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮 编 100190
电子信箱 jig@aircas.ac.cn
电 话 010-58887035
网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订 购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
(邮政信箱: 北京399信箱 邮编: 100048)
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

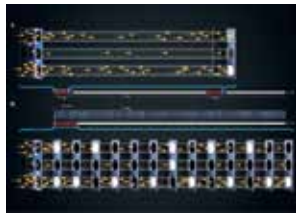
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元



面向智慧交通的图像处理与
边缘计算(第1743页)



脉冲视觉研究进展(第1823
页)



移动在线实时绘制技术研究
综述(第1877页)

序言王耀南 I

视觉理解与计算成像

基于深度学习的视觉目标检测技术综述

曹家乐, 李亚利, 孙汉卿, 谢今, 黄凯奇, 庞彦伟 1697

面向复杂场景的人物视觉理解技术

马利庄, 吴飞, 毛启容, 王鹏杰, 陈玉珑 1723

面向智慧交通的图像处理与边缘计算

曹行健, 张志涛, 孙彦赞, 王平, 徐树公, 刘富强, 王超, 彭飞, 穆世义, 刘文予, 杨铀 ... 1743

视觉弱监督学习研究进展

任冬伟, 王旗龙, 魏云超, 孟德宇, 左旺孟 1768

智能遥感: AI赋能遥感技术

孙显, 孟瑜, 刁文辉, 黄丽佳, 张新, 骆剑承, 高连如, 王佩瑾, 闫志远, 郜丽静, 董文, 冯瑛
超, 李霁豪, 付琨 1799

脉冲视觉研究进展

黄铁军, 余肇飞, 李源, 施柏鑫, 熊瑞勤, 马雷, 王威 1823

计算成像前沿进展

顿雄, 付强, 李浩天, 孙天成, 王建, 孙启霖 1840

移动在线实时绘制技术研究综述

刘畅, 霍宇驰, 张严辞, 张乾, 郑家祥, 唐睿, 余耿, 王锐, 贾金原 1877

数据挖掘和信息交互

表格识别技术研究进展

高良才, 李一博, 都林, 张新鹏, 朱子仪, 卢宁, 金连文, 黄永帅, 汤帆 1898

多媒体隐写研究进展

张卫明, 王宏霞, 李斌, 任延珍, 杨忠良, 陈可江, 李伟祥, 张新鹏, 俞能海 1918

大脑多模态成像技术定量研究进展

叶慧慧, 何宏建, 方静苑, 童琪琦, 周子涵, 刘华锋 1944

多模态人机交互综述

陶建华, 巫英才, 喻纯, 翁冬冬, 李冠君, 韩腾, 王运涛, 刘斌 1956

文化遗产活化关键技术研究进展

耿国华, 何雪磊, 王美丽, 李康, 贺小伟 1988

情感计算与理解研究发展概述

姚鸿勋, 邓伟洪, 刘洪海, 洪晓鹏, 王甦菁, 杨巨峰, 赵思成 2008

跨模态脑图谱数据融合研究进展

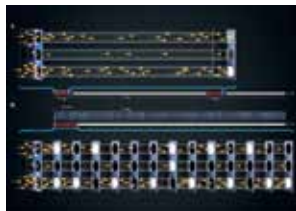
罗娜, 宋明, 杨正宜, 蒋田仔 2036

CONTENTS

JOURNAL OF IMAGE AND GRAPHICS



The review of image processing and edge computing for intelligent transportation system(P1743)



Advances in spike vision(P1823)



A review of real-time rendering technology based on mobile internet platforms(P1877)

Visual Understanding & Computational Imaging

A survey on deep learning based visual object detection

Cao Jiale, Li Yali, Sun Hanqing, Xie Jin, Huang Kaiqi, Pang Yanwei 1697

Visual recognition technologies for complex scenarios analysis

Ma Lizhuang, Wu Fei, Mao Qirong, Wang Pengjie, Chen Yulong 1723

The review of image processing and edge computing for intelligent transportation system

Cao Xingjian, Zhang Zhitao, Sun Yanzan, Wang Ping, Xu Shugong, Liu Fuqiang, Wang Chao, Peng Fei, Mu Shiyi, Liu Wenyu, Yang You 1743

Progress in weakly supervised learning for visual understanding

Ren Dongwei, Wang Qilong, Wei Yunchao, Meng Deyu, Zuo Wangmeng 1768

The review of AI-based intelligent remote sensing capabilities

Sun Xian, Meng Yu, Diao Wenhui, Huang Lijia, Zhang Xin, Luo Jiancheng, Gao Lianru, Wang Peijin, Yan Zhiyuan, Gao Lijing, Dong Wen, Feng Yingchao, Li Jihao, Fu Kun 1799

Advances in spike vision

Huang Tiejun, Yu Zhaofei, Li Yuan, Shi Boxin, Xiong Ruiqin, Ma Lei, Wang Wei 1823

Recent progress in computational imaging

Dun Xiong, Fu Qiang, Li Haotian, Sun Tiancheng, Wang Jian, Sun Qilin 1840

A review of real-time rendering technology based on mobile internet platforms

Liu Chang, Huo Yuchi, Zhang Yanchi, Zhang Qian, Zheng Jiaxiang, Tang Rui, Yu Geng, Wang Rui, Jia Jinyuan 1877

Data Mining and Information Interaction

A survey on table recognition technology

Gao Liangcai, Li Yibo, Du Lin, Zhang Xinpeng, Zhu Ziyi, Lu Ning, Jin Lianwen, Huang Yongshuai, Tang Zhi 1898

Overview of steganography on multimedia

Zhang Weiming, Wang Hongxia, Li Bin, Ren Yanzhen, Yang Zhongliang, Chen Kejiang, Li Weixiang, Zhang Xinpeng, Yu Nenghai 1918

Research progress of quantitative multimodal brain imaging technology

Ye Huihui, He Hongjian, Fang Jingwan, Tong Qiqi, Zhou Zihan, Liu Huafeng 1944

A survey on multi-modal human-computer interaction

Tao Jianhua, Wu Yingcai, Yu Chun, Weng Dongdong, Li Guanjuan, Han Teng, Wang Yuntao, Liu Bin 1956

Research progress on key technologies of cultural heritage activation

Geng Guohua, He Xuelei, Wang Meili, Li Kang, He Xiaowei 1988

An overview of research development of affective computing and understanding

Yao Hongxun, Deng Weihong, Liu Honghai, Hong Xiaopeng, Wang Sujing, Yang Jufeng, Zhao Sicheng 2008

Survey on cross-modal fusion based on brain atlas data

Luo Na, Song Ming, Yang Zhengyi, Jiang Tianzi 2036

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2022)06-1697-26

论文引用格式: Cao J L, Li Y L, Sun H Q, Xie J, Huang K Q and Pang Y W. 2022. A survey on deep learning based visual object detection. Journal of Image and Graphics, 27(06): 1697-1722 (曹家乐, 李亚利, 孙汉卿, 谢今, 黄凯奇, 庞彦伟. 2022. 基于深度学习的视觉目标检测技术综述. 中国图象图形学报, 27(06): 1697-1722) [DOI:10.11834/jig.220069]

基于深度学习的视觉目标检测技术综述

曹家乐¹, 李亚利², 孙汉卿¹, 谢今³, 黄凯奇⁴, 庞彦伟^{1*}

1. 天津大学, 天津 300072; 2. 清华大学, 北京 100084;

3. 重庆大学, 重庆 400044; 4. 中国科学院自动化研究所, 北京 100190

摘要: 视觉目标检测旨在定位和识别图像中存在的物体, 属于计算机视觉领域的经典任务之一, 也是许多计算机视觉任务的前提与基础, 在自动驾驶、视频监控等领域具有重要的应用价值, 受到研究人员的广泛关注。随着深度学习技术的飞速发展, 目标检测取得了巨大的进展。首先, 本文总结了深度目标检测在训练和测试过程中的基本流程。训练阶段包括数据预处理、检测网络、标签分配与损失函数计算等过程, 测试阶段使用经过训练的检测器生成检测结果并对检测结果进行后处理。然后, 回顾基于单目相机的视觉目标检测方法, 主要包括基于锚点框的方法、无锚点框的方法和端到端预测的方法等。同时, 总结了目标检测中一些常见的子模块设计方法。在基于单目相机的视觉目标检测方法之后, 介绍了基于双目相机的视觉目标检测方法。在此基础上, 分别对比了单目标检测和双目目标检测的国内外研究进展情况, 并展望了视觉目标检测技术发展趋势。通过总结和分析, 希望能够为相关研究人员进行视觉目标检测相关研究提供参考。

关键词: 视觉目标检测; 深度学习; 单目; 双目; 锚点框

A survey on deep learning based visual object detection

Cao Jiale¹, Li Yali², Sun Hanqing¹, Xie Jin³, Huang Kaiqi⁴, Pang Yanwei^{1*}

1. Tianjin University, Tianjin 300072, China; 2. Tsinghua University, Beijing 100084, China;

3. Chongqing University, Chongqing 400044, China; 4. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Visual object detection aims to locate and recognize objects in images, which is one of the classical tasks in the field of computer vision and also the premise and foundation of many computer vision tasks. Visual object detection plays a very important role in the applications of automatic driving, video surveillance, which has attracted extensive attention of the researches in past few decades. In recent years, with the development of the technique of deep learning, visual object detection has also made great progress. This paper focuses on a deep survey on deep learning based visual object detection, including monocular object detection and stereo object detection. First, we summarize the pipeline of deep object detection during the training and inference. The training process is composed of data pre-processing, detection network design, and label assignment and loss function in common. Data pre-processing (e. g., multi-scale training and flip) aims to enhance the diversity of the given training data, which can improve detection performance of object detector. Detection network usually consists of three key parts like the backbone (e. g., Visual Geometry Group (VGG) and ResNet), feature fusion

收稿日期: 2022-01-18; 修回日期: 2022-02-25; 预印本日期: 2022-03-04

* 通信作者: 庞彦伟 pyw@tju.edu.cn

基金项目: 国家重点研发计划资助(2018AAA0102800); 中国博士后科学基金项目(2021M700613); 中国人工智能学会—华为 MindSpore 学术奖励基金

Supported by: National Key R&D Program of China (2018AAA0102800); China Postdoctoral Science Foundation (2021M700613)

module (e.g., feature pyramid network (FPN)), and prediction network (e.g., region of interest head network (RoI head)). Label assignment aims to assign the true value for each prediction, and loss function can supervise the network training. During inference, we adopt the trained detector to generate the detection bounding-boxes and employ the post-processing step (e.g., non-maximum suppression (NMS)) to combine the bounding-boxes. Second, we illustrate a deep review on monocular object detection, including anchor-based, anchor-free, and end-to-end methods, respectively. Anchor-based methods design some default anchors and perform classification and regression based on these default anchors, which can be further split into two-stage and one-stage methods. Two-stage methods first generate some candidate proposals based on the default anchors, and second classify/regress these proposals. Compared to two-stage methods, one-stage methods directly perform classification and regression on default anchors directly, which usually have a faster inference speed. The representative two-stage methods are regional-based convolutional neural network (R-CNN) series, and the representative one-stage methods are you only look once (YOLO) and single shot detector (SSD). Compared to anchor-based methods, more robust anchor-free methods perform classification and regression without any hand-crafted default anchors. We split anchor-free methods into keypoint-based methods, and center-based methods. Keypoint-based methods predict multiple keypoints of objects for localization, while center-based methods predict the left, right, top, and bottom distances to the object boundary. The representative keypoint-based method is CornerNet, and the representative center-based methods are fully convolutional one-stage detector (FCOS) and CenterNet. The anchor-based methods and anchor-free methods mentioned above require the post-processing to remove the redundant detection results for each object in common. To solve this problem, the recently introduced end-to-end methods directly predict one bounding box for each object straightforward, which can avoid the post-processing. The representative end-to-end method is detection transformer (DETR) that performs prediction via a transformer module. In addition, we review some classical modules employed in monocular object detection, including feature pyramid structure, prediction network design, label assignment and loss function. The feature pyramid structure employs different layers to detect multi-scaled objects, which can deal with scale variance issue. Prediction network design contains the re-designs of classification and regression, which aims to better deal with these two sub-tasks. Label assignment and loss function aim to better guide detector training. Third, we introduce stereo object detection. According to the coordinate space of features, existing detectors are divided into two categories: frustum-based and inverse-projection-based approaches. Frustum-based approaches directly predict 3D objects on features in the image frustum space. Stereo R-CNN and StereoCenterNet construct stereo features in the image frustum space via concatenating a pair of unary features concatenation. Plane-sweeping is another method of constructing frustum features as cost volumes, which is used in instance-depth-aware 3D detection (IDA-3D) and YOLOStereo3D. In contrast to the frustum-based approaches, inverse-projection-based approaches explicitly project the pixels or frustum features into 3D Cartesian space. There are mainly three manners of the inverse projection: projecting all pixels to the full 3D space as a pseudo point cloud, projecting the cost volume features to 3D feature volume features, or projecting the pixels in each region proposal to an instance-level point cloud. Pseudo-LiDAR is a pioneer method that transforms stereo images to their point cloud representation, which embraces the advances in both disparity estimation and LiDAR-based 3D detection. Deep stereo geometry network (DSGN) projects the frustum-based cost volume features to 3D volume features and further squeezes them into bird's eye view (BEV) for detection. Disp R-CNN leverages Mask R-CNN, a representative 2D instance segmentation model, and generates a set of instance-level point clouds for each stereo image pair. Based on the summary of monocular and stereo object detection, we further compare the progress of domestic and foreign researches, and present some representative universities or companies on visual object detection. Finally, we present some development tendency in visual object detection, including efficient end-to-end object detection, self-supervised object detection, long-tailed object detection, few-shot and zero-shot object detection, large-scale stereo object detection dataset, weakly-supervised stereo object detection.

Key words: visual object detection; deep learning; monocular; stereo; anchor

0 引言

视觉目标检测是计算机视觉的经典任务,旨在定位图像中存在物体的位置并识别物体的具体类别。目标检测是许多计算机视觉任务及相关应用的基础与前提,直接决定相关视觉任务及应用的性能好坏。因此,视觉目标检测技术受到了学术界、工业界等各领域、乃至世界各国政府的广泛关注。在学术界,目标检测一直是各大计算机视觉会议及期刊的研究热点之一,每年有大量的目标检测相关论文发表。根据谷歌学术显示,研究人员近10年来在目标检测方面发表论文15 000余篇。在工业界,国内外科技巨头(如谷歌、脸书、华为和百度等)、初创公司(如商汤、旷视等)纷纷在目标检测相关领域投入大量人力财力。与此同时,目标检测技术是新一代人工智能的重要共性关键技术,世界各国竞相竞争。

在过去的几十年中,目标检测经历了基于手工设计特征的方法到基于深度特征的方法等不同发展阶段。早期,目标检测方法通常采用手工设计特征加浅层分类器的技术路线,例如支持向量机(support vector machines, SVM)和AdaBoost等,涌现了包括Haar特征(Viola和Jones, 2004)、方向梯度直方图(histograms of oriented gradients, HOG)特征(Dalal和Triggs, 2005)等一系列经典的目标检测特征描述子。2012年以来,深度学习技术取得了飞速的发展,并行计算资源不断迭代更新,大规模数据库及评测标准相继构建与公开。基于上述技术、算力和数据的铺垫,视觉目标检测开始在精度与效率等方面取得了显著的进展,先后涌现出区域卷积神经网络(region-based convolutional neural network, R-CNN)(Girshick等, 2014)、SSD(single shot detector)(Liu等, 2016)、YOLO(you only look once)(Redmon等, 2016)、DETR(detection transformer)(Carion等, 2020)等一系列经典的研究工作。相比于传统手工设计特征的方法,基于深度学习的方法避免了烦琐的手工设计过程,能够自动学习更具有区分力的深度特征。与此同时,基于深度学习的方法将特征提取和分类器学习统一在一个框架中,能够进行端到端的学习。

随着技术的不断发展与成熟,深度目标检测技术开始在实际应用中发挥重要作用。近些年,国内

外涌现了一批以目标检测等视觉技术为核心技术的科技创业公司,如旷视科技、商汤科技等。同时,视觉目标检测是自动驾驶汽车环境感知重要的内容之一,以特斯拉为代表的一批科技公司甚至采用纯视觉目标感知的技术路线开展自动驾驶研究。尽管目标检测技术已经开始走向实际应用,但是当前目标检测的性能仍然无法到达人类视觉的性能,存在巨大改进与提升的空间。

鉴于基于深度学习的目标检测技术在学术界和产业界取得了巨大成功,本文对基于深度学习的视觉目标检测技术进行了系统的总结和分析,包括国内外研究现状以及未来的发展趋势等。根据视觉目标检测采用视觉传感器的数量不同,将视觉目标检测分为两类:基于单目相机的视觉目标检测和基于双目相机的视觉目标检测。相比于单目相机,双目相机能够提供3维信息。因此,基于双目相机的视觉目标检测能够提供精准的目标3维信息,在自动驾驶等领域能够更好地满足应用需求。

首先介绍目标检测的基本流程,包括训练和测试过程。接着,系统地总结和分析单目视觉目标检测。然后,介绍双目视觉目标检测。最终,对比国内外发展现状,并对发展趋势进行展望。

1 基于深度学习的目标检测基本流程

如图1所示,基于深度学习的目标检测主要包括训练和测试两个部分。训练的主要目的是利用训练数据集进行检测网络的参数学习。训练数据集包含大量的视觉图像及标注信息(物体位置及类别)。如图1(a)所示,训练阶段的主要过程包括数据预处理、检测网络以及标签匹配与损失计算等部分。

1)数据预处理。数据预处理旨在增强训练数据多样性,进而提升检测网络的检测能力。常用的数据增强手段有翻转、缩放、均值归一化和色调变化等。除此之外,研究人员在数据预处理方面做了大量的研究工作。一些研究人员提出从图像中擦除部分子区域,如CutOut(DeVries和Taylor, 2017)、Random erasing(Zhong等, 2020b)、HaS(hide-and-seek)(Singh和Lee, 2017)、GridMask(Chen等, 2020a)等。Zhang等人(2018a)通过将不同图像和标签进行差值表示提升分类性能,简称为MixUp。Yun等人(2019)认为直接擦除图像子区域会造成信息损失,

提出将其他训练图像粘贴到擦除的子区域,简称为 CutMix。类似地,Fang 等人(2019)将其他图像的实例掩膜粘贴到当前图像用于实例分割。此外,研究人员提出将多个图像拼接在一起进行训练,提升检

测器应对尺度变化的鲁棒性,如 Mosaic(Bochkovskiy 等,2020)、Montage(Zhou 等,2020)、DST(dynamic scale training)(Chen 等,2020d)。此后,Chen 等人(2021e)提出自动搜索的尺度增强策略。

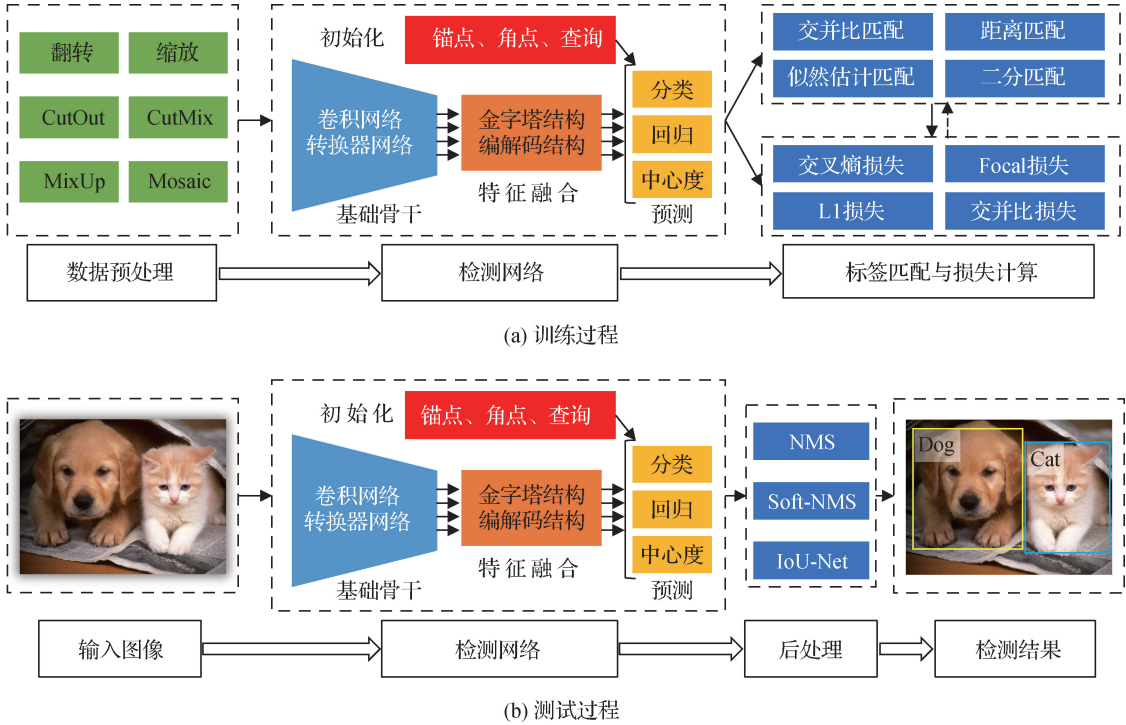


图1 基于深度学习的视觉目标检测训练与测试

Fig. 1 Training and inference of deep learning based visual object detection ((a) training stage; (b) test stage)

2)检测网络。检测网络一般包括基础骨干、特征融合及预测网络3部分。目标检测器的基础骨干通常采用用于图像分类的深度卷积网络,如 AlexNet(Krizhevsky 等,2012)、VGGNet(Visual Geometry Group)(Simonyan 和 Zisserman,2014)、ResNet(He 等,2016)和 DenseNet(Huang 等,2017)等。近期,研究人员开始采用基于 Transformer(Vaswani 等,2017)的基础骨干网络,如 ViT(vision transformer)(Dosovitskiy 等,2021;Beal 等,2020)、Swin(Liu 等,2021c)和 PVT(pyramid vision transformer)(Wang 等,2021c)等。通常将大规模图像分类数据库 ImageNet(Russakovsky 等,2015)([https://www. image-net. org/](https://www.image-net.org/))上的预训练权重作为检测器骨干网络的初始权重。特征融合主要是对基础骨干提取的特征进行融合,用于后续分类和回归。常见的特征融合方式是特征金字塔结构(Lin 等,2017a)。研究人员开始用基于 Transformer 编解码的特征融合方式进行目标检测。最后,预测网络进行分类和回归等任

务。在两阶段目标检测方法中,分类和回归通常采用全连接的方式,而在单阶段的方法中,分类和回归等通常采用全卷积的方式。Guo 等人(2020b)利用神经网络搜索技术同时搜索基础骨干、特征融合和预测网络等3部分。与此同时,检测器通常还需要一些初始化,如锚点框初始化、角点初始化和查询特征初始化等。

3)标签分配与损失计算。标签分配主要是为检测器预测提供真实值。在目标检测中,标签分配的准则包括交并比(intersection over union, IoU)准则、距离准则、似然估计准则和二分匹配等。交并比准则通常用于基于锚点框的目标检测方法,根据锚点框与物体真实框之间的交并比将锚点框分配到对应的物体。距离准则通常用于无锚点框的目标检测方法,根据点到物体中心的距离将其分配到对应的物体。似然估计准则和二分匹配通常基于分类和回归的联合损失进行最优标签分配。基于标签分类的结果,采用损失函数计算分类和回归等任务的损失,

并利用反向传播算法更新检测网络的权重。常用的分类损失函数有交叉熵损失函数、聚焦损失函数(Lin等,2017b)等,而回归损失函数有L1损失函数、平滑L1损失函数、交并比IoU损失函数、GIoU(generalized IoU)损失函数(Rezatofighi等,2019)和CIoU(complete-IoU)损失函数(Zheng等,2020b)等。

基于训练阶段学习的检测网络,在测试阶段输出给定图像中存在物体的类别以及位置信息。如图1(b)所示,主要包括输入图像、检测网络和后处理等过程。对于一幅给定的图像,先利用训练好的检测网络生成分类和回归结果。一般而言,大部分目标检测方法在同一物体周围会生成多个检测结果。因此,大部分目标检测方法需要一个后处理步骤,旨在为每个物体保留一个检测结果并去除其他冗余的检测结果。最常用的后处理方法为非极大值抑制方法(non-maximum suppression, NMS)。NMS试图为每个物体保留一个分类得分最高的检测结果。Bodla等人(2017)认为NMS方法容易将距离较近的多个物体检测结果合并,造成部分物体漏检的问题。为解决这一问题,Bodla等人(2017)对NMS进行改进并提出Soft-NMS。该方法通过降低交并比高的检测结果的分类得分来抑制冗余检测。Jiang等人(2018)提出IoUNet,预测检测框与物体真实框之间的交并比,并根据预测的交并比值进行非极大值抑制。He等人(2018)提出学习检测框的定位方差,并利用定位方差线性加权邻近检测框来提升当前检测框的定位精度。Pato等人(2020)通过对检测结果上下文推理实现对检测结果的重打分。

视觉目标检测在训练和测试过程相对烦琐。为了更好地促进目标检测技术的发展,方便不同方法进行公平比较,国内外研究人员先后发布了不同的目标检测开源平台,使用基于模块化设计的思想,方便支持不同目标检测方法的集成。国外比较有代表性的研究机构是美国FaceBook人工智能研究院,先后发布了Detectron(<https://github.com/facebookresearch/Detectron>)、maskrcnn-benchmark(<https://github.com/facebookresearch/maskrcnn-benchmark>)和Detectron2(<https://github.com/facebookresearch/detectron2>)等目标检测与分割开源平台。国内比较有代表性的机构是商汤科技和图森科技,发布了mmdetection(<https://github.com/open-mmlab/mmdetection>)和SimpleDet(<https://github.com/TuSimple/simpliedet>)等

目标检测开源平台。上述目标检测平台大多基于国外深度学习核心架构Caffe2(<https://github.com/facebookarchive/caffe2>)、PyTorch(<https://pytorch.org/>)和MXNet(<https://mxnet.apache.org/versions/1.8.0/>)。与此同时,百度、华为和清华大学等国内科技公司与大学相继发布了深度学习核心架构PaddlePaddle(<https://www.paddlepaddle.org.cn/>)、MindSpore(<https://www.mindspore.cn/>)和Jittor(<https://cg.cs.tsinghua.edu.cn/jittor/>)等,并提供了一些典型的目标检测方法接口,促进目标检测技术走向实际应用。

2 单目视觉目标检测

单目视觉目标检测是视觉目标检测的基础,旨在预测单幅图像中存在物体的位置以及类别信息。自2012年深度卷积神经网络在图像分类任务取得成功(Krizhevsky等,2012),研究人员开始尝试用深度卷积神经网络进行目标检测,如DetectorNet(Szegedy等,2013)和OcarFeat(Sermanet等,2014)。此后,基于深度学习的目标检测开始主导目标检测的发展。图2给出了近年基于深度学习的目标检测发展历程,并列出了一些具有代表性的深度网络模型及目标检测方法。灰色字体表示一些代表性的深度网络模型,黑色字体表示一些代表性的深度学习目标检测方法。将目标检测方法分为3类:基于锚点框的目标检测方法、无锚点框的目标检测方法以及端到端预测的目标检测方法。需要指出的是,端到端预测的目标检测方法属于无锚点框的目标检测方法。由于端到端预测的目标检测方法不需要后处理操作,大多采用转换器模型直接为每个目标预测一个检测结果,是一个更简洁的检测架构,将其单独归为一类进行详细介绍。

2.1 基于锚点框的目标检测方法

基于锚点框的目标检测方法为空间每一个位置设定多个矩形框,以便尽可能地覆盖图像中所有存在的物体。基于锚点框的目标检测可以分为两类(赵永强等,2020):两阶段目标检测方法和单阶段目标检测方法。图3给出两类方法的基本架构图。两阶段方法(图3(a))首先提取 k 个类别不具体的候选检测窗口,然后进一步对这些候选检测窗口进行分类和回归,生成最终的检测结果。与两阶段方法

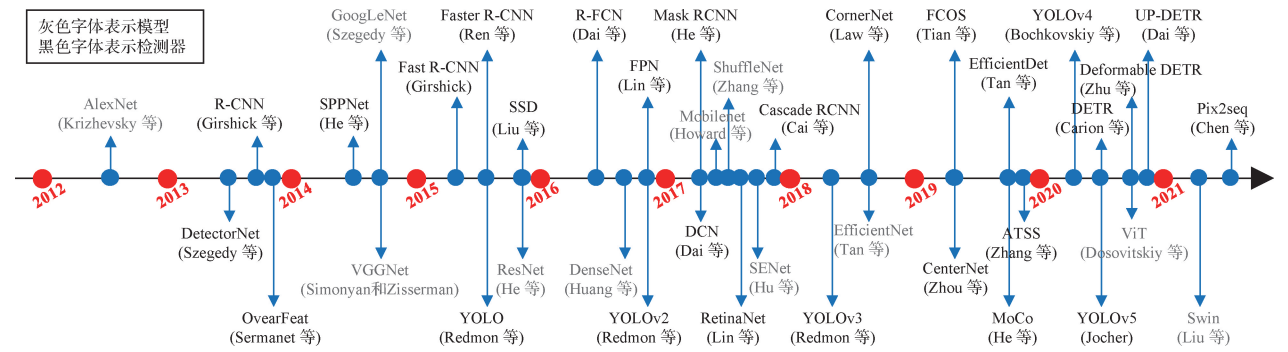


图2 基于深度学习的视觉目标检测发展过程(图中括号内为作者信息)

Fig. 2 Development process of visual object detection based on deep learning

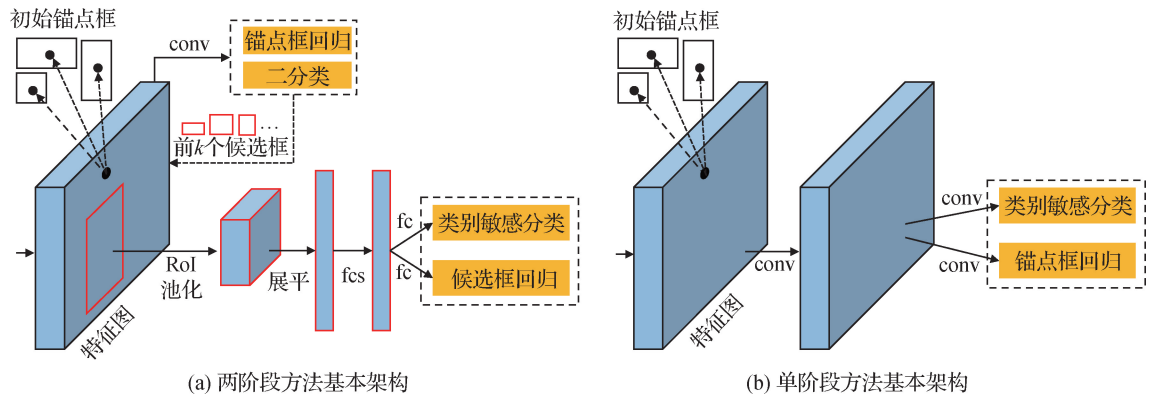


图3 基于锚点框的目标检测方法基本架构

Fig. 3 Architectures of anchor-based object detection methods ((a) two-stage architecture; (b) one-stage architecture)

不同,单阶段方法(图3(b))直接对锚点框进行分类和回归。一般而言,两阶段方法具有较高的检测精度,而单阶段方法具有较快的推理速度。

2.1.1 两阶段目标检测方法

区域卷积神经网络(region-based convolutional neural network, R-CNN)系列工作是两阶段目标检测方法的最主要代表性工作。R-CNN(Girshick等, 2014)首先利用选择性搜索方法(Uijlings等, 2013)生成2 000个最可能是物体的候选检测框,然后利用深度卷积神经网络提取这些候选检测框的深度特征,最后利用支持向量机进行分类和回归。该方法在当时取得了巨大的成功,大幅度提升了目标检测的精度。由于R-CNN分别提取每个候选框的深度特征,因此存在推理速度慢的问题。针对这一问题, He等人(2015)先通过特征共享的方式提取整幅图像的特征,然后利用空间金字塔池化(spatial pyramid pooling, SPP)操作,将每个候选框对应的特征转换成固定长度的特征,进行后续SVM的分类和回

归,该方法简称为SPPNet。Girshick(2015)认为R-CNN和SPPNet特征提取和预测(分类和回归)是一个多阶段的过程,限制了深度神经网络的性能,并提出了R-CNN的改进工作Fast R-CNN。Fast R-CNN首先提取整幅图像的深度特征,然后利用感兴趣区域(region of interest, RoI)池化操作将候选检测框的特征缩放至固定大小,最终利用全连接层进行分类和回归。由于感兴趣池化操作能够实现反向传播, Fast R-CNN能够联合训练整个网络。随后, Ren等人(2015)提出Faster R-CNN,进一步将候选窗口的生成同候选窗口的分类与回归统一到在一个网络中联合学习。

在Faster R-CNN的基础上,研究人员进行了大量的改进。一些研究人员关注RoI池化操作。Dai等人(2016)提出位置敏感感兴趣区域(position-sensitive RoI, PSRoI)池化操作,从特征图的不同通道累积空间对应位置的特征。Zhu等人(2017)认为RoI池化操作可以提取上下文信息, PSRoI能够捕获物

体的局部信息。基于此假设,Zhu 等人(2017)提出了 CoupleNet,将 RoI 池化操作提取的特征和 PSRoI 池化操作提取的特征进行融合,用于后续的分类和回归。Dai 等人(2017)提出了可变形 RoI 池化操作,能够更好地刻画物体的形变。He 等人(2017)提出 RoIAlign 池化操作,解决 RoI 池化操作因量化误差带来的特征不匹配问题。此外,一些研究人员关注级联结构在两阶段方法中的应用。Cai 和 Vasconcelos(2018)提出级联目标检测架构 Cascade R-CNN,将多个 Fast R-CNN 头网络级联起来,当前级对前一级的分类和回归结果进一步进行分类和回归。类似地,Zhong 等人(2020a)和 Vu 等人(2019)将级联思想用于候选窗口生成。

为了应对物体尺度的变化,研究人员提出了基于图像金字塔的方法(Singh 和 Davis,2018; Singh 等,2018)和基于特征金字塔的方法(Lin 等,2017a)。基于图像金字塔的方法采用不同尺度的图像检测不同尺度的物体,如小尺度图像检测大尺度物体、大尺度图像检测小尺度物体。基于图像金字塔的方法需要利用检测网络分别检测多个不同尺度的图像,计算量相对较大。基于特征金字塔的方法采用单个检测网络内部不同层检测不同尺度的物体,计算量相对较少(李晖等,2020;姜文涛等,2019)。因此,研究人员更多关注基于特征金字塔的方法(姚群力等,2019)。

2.1.2 单阶段目标检测方法

YOLO(you only look once)系列工作是单阶段目标检测方法的代表性工作之一。YOLO(Redmon 等,2016)直接将图像分成 $N \times N$ 大小的子区域,并预测每个子区域存在物体的概率、类别以及位置偏移量。YOLO 结构十分简单,具有很快的运算速度。此后,YOLOv2(Redmon 和 Farhadi,2017)、YOLOv3(Redmon 和 Farhadi,2018)、YOLOv4(Bochkovskiy 等,2020)和 YOLOv5(<https://github.com/ultralytics/yolov5>)等相继提出,获得了广泛的关注(张伟等,2021)。YOLOv2 引入包括批归一化操作、高分辨率输入和全卷积操作等改进,使其能够在保持较快检测速度的情况下提升目标检测精度。YOLOv3 提出 Darknet-53 基础骨干网络和多尺度预测等改进。YOLOv4 对数据预处理、检测网络设计和预测网络等过程进行系统的分析,并基于这些分析设计

了适合单显卡的高效目标检测器。YOLOv5 提供 4 种不同大小的目标检测器,以便满足不同应用的需求。

SSD(single shot detector)(Liu 等,2016)是另一个代表性的单阶段目标检测方法。为了检测不同尺度的物体,SSD 采用不同层的特征图检测不同尺度的物体。靠前分辨率高的特征图检测小尺度物体,靠后分辨率低的特征图检测大尺度物体。此后,研究人员在 SSD 的基础上开展了大量的工作。Fu 等人(2017)和 Kong 等人(2018)通过去卷积操作作为 SSD 引入上下文信息。Zhou 等人(2018)提出尺度转换层,将高语义特征图转换成不同尺度的特征图检测不同尺度的物体,保持了特征图的语义一致性。受人类视觉系统启发,Liu 等人(2018a)提出利用不同膨胀率的卷积层提取不同感受野的上下文信息。Zhao 等人(2019)、Kim 等人(2018)和 Wang 等人(2019b)分别提出多级结构、并行结构和图像金字塔结构增加特征金字塔结构的上下文信息。Zhang 等人(2018b)、Cao 等人(2019b)和 Nie 等人(2019)提出了基于级联结构的单阶段目标检测方法,提高目标检测定位精度。Zhang 等人(2018c)、Dvornik 等人(2017)和 Cao 等人(2019a)联合检测和分割两个任务,试图提升多任务学习的性能。Li 等人(2020d)提出了尺度解混模块,使得不同层能够更好地关注不同尺度的物体。为了解决类别不平衡问题,Lin 等人(2017b)提出了聚焦损失(focal loss)函数,在训练过程中增大难样本的损失权重。

2.2 无锚点框目标检测方法

基于锚点框的目标检测方法需要人为地根据数据库特性设定锚点框的尺度和长宽比等参数。因而,基于锚点框的目标检测方法存在检测器对参数敏感、检测器泛化能力差等问题。针对这一问题,研究人员提出了无锚点框的目标检测方法。无锚点框目标检测方法主要分为两类:基于关键点的目标检测方法和基于内部点的目标检测方法。图 4(a)给出了基于关键点方法的基本结构图。基于关键点的方法通常通过预测物体的多个关键点,并将关键点集成实现对物体的检测。图 4(b)给出了基于内部点方法的基本结构图。基于内部点的方法预测物体内部点到物体边界的上下左右偏移量及内部点所属的类别信息等。

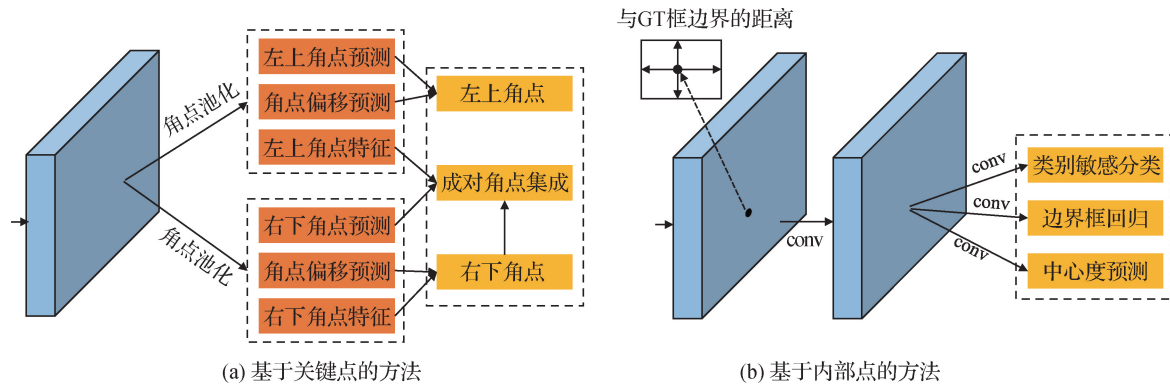


图4 无锚点框目标检测方法基本架构

Fig. 4 Architectures of anchor-free object detection methods ((a) keypoint-based method; (b) center-based method)

2.2.1 基于关键点的目标检测方法

2018年, Law和Deng(2018)创新性地提出了基于角点的目标检测方法 CornerNet。基于全卷积神经网络 Hourglass(Newell等, 2016)输出的高分辨率特征图, CornerNet 分别预测物体左上角点热图、右下角点热图以及两个角点的集成特征向量。基于集成特征的相似性度, CornerNet 将属于同一物体的两个角点关联起来构成一个物体。此外, CornerNet 预测类别敏感的角点热图来区分不同类别的物体, 并提出角点池化层提升网络对角点的检测能力。该方法消除了单阶段检测方法需要设置锚点框的需求。为提升 CornerNet 推理有效性, Law等人(2020)从减少处理像素数及每个像素的运算量等两方面出发, 提出了 CornerNet 的快速版本 CornerNet-Lite。CornerNet-Lite 主要包括两个模块: CornerNet-Saccade 模块和 CornerNet-Squeeze 模块。CornerNet-Saccade 从低分辨率输入图像中快速预测可能存在物体的候选区域, 而 CornerNet-Squeeze 采用轻量卷积神经网络从高分辨率候选区域中精准检测物体。

基于 CornerNet 的思想, 研究人员进行了一些改进工作。Zhou等人(2019b)认为物体的角点通常位于物体外, 缺少物体外观特征信息, 为了解决这一问题, 提出了基于极值点的目标检测器 ExtremeNet, 预测4个类别敏感的极值点热图和1个类别敏感的中心点热图, 并通过判断4个极值点对应的中心点响应值是否大于预定阈值来集成极值点。4个极值点分别是左极值点、右极值点、上极值点和下极值点。为了能够获取更多物体外观信息, 减少 CornerNet 生成的大量虚检点, Duan等人(2019)在 CornerNet 的基础上增加对中心点的预测。类似地, 为了减少对

角点错误匹配的数量, Dong等人(2020)在预测成对角点的同时预测成对角点的向心偏移量, 并根据它们是否都靠近中心来判断是否属于同一物体。Duan等人(2020)利用角点目标检测提取网络提出候选检测框, 进而利用 Fast R-CNN 头网络对这些候选框进行分类和回归。

上述方法都需要将不同关键点集成构成一个物体。Yang等人(2019)直接预测一组关键点表示物体, 并利用这组关键点所对应的特征对物体进行分类, 简称 RepPoints。与 CornerNet 和 ExtremeNet 等自下而上的方法比, RepPoints 不需要将不同关键点集成并采用了更准确的特征进行分类。在此基础上, Yang等人(2020)提出基于稠密关键点的 RePoints, 用于更精准的实例分割任务。Chen等人(2020e)通过引入两个辅助任务(即角点预测和前景预测)增强 RepPoints 提取的特征和微调 RepPoints 的检测框定位, 进而提升 RepPoints 的物体检测准确率。Wei等人(2020)利用若干个初始关键点表示物体, 基于这些初始关键点对物体进行回归和分类。

2.2.2 基于内部点的目标检测方法

2015年起, 研究人员已经提出了基于内部点的目标检测方法(如 DenseBox(Huang等, 2015)和 UnitBox(Yu等, 2016))。但是, 这些方法仅用于人脸检测等单一类别视觉目标检测任务上。自2019年开始, 研究人员将基于内部点的方法用于一般目标检测任务中。Zhu等人(2019)率先提出特征选择性的无锚点目标检测方法 FSAF(feature selective anchor-free), 预测物体中心区域到物体边界的偏移量, 并根据分类和回归损失动态地将物体分配到最优的金字塔尺度上预测。Tian等人(2019)提出目

标检测方法 FCOS (fully convolutional one-stage detector), 预测物体所有内部点到其上下左右边界的距离及物体的类别。为了检测不同尺度的物体, FCOS 根据物体的尺度将其分配到金字塔结构的不同层进行预测。同一时期, Kong 等人(2019)提出无锚点框检测方法 FoveaBox, 仅利用部分中心区域点预测物体。Wang 等人(2020b)提出基于网络架构搜索(neural architecture search, NAS)的检测方法 NAS-FCOS, 采用网络架构搜索的思想构建特征金字塔结构和预测头网络。Qiu 等人(2020b)提出基于边界强化模块的目标检测方法 BorderDet, 提取 FCOS 输出提取边界框上局部最大特征值进行第2次分类和回归, 巩固第1次分类和回归结果。Wang 等人(2019a)提出预测物体的位置以及尺度等信息生成更聚焦在物体周围的锚点框, 进而提取更好的候选窗口用于后续分类和回归。

与此同时, 研究人员提出了快速高效的检测方法。Zhou 等人(2019a)提出高效的无锚点框检测方法 CenterNet, 预测物体的中心点、中心点偏移量以及物体的宽高信息等。同时, CenterNet 采用最大池化操作提取峰值点, 避免了非极大值抑制后处理操作。Liu 等人(2020b)改进高斯核对训练样本进行编码的方式, 加快了 CenterNet 的训练速度。Lan 等人(2020)在 CenterNet 的基础上增加角点学习监督, 增强了对物体边界特征的学习。

此外, Qiu 等人(2021)提出了基于十字线表示的目标检测方法 CrossDet, 采用一对交叉的水平线和竖直线表示物体。为了能够更好地捕获交叉线上的特征用于分类、水平回归以及竖直回归, CrossDet 提出了基于水平池化和竖直池化的交叉线特征提取模块。

2.3 端对端预测的目标检测方法

上述基于锚点框的目标检测方法和无锚点框的目标检测方法通常存在一个物体对多个检测框的情况。实际中, 同一物体只需要保留一个检测框即可。因此, 上述方法一般都需要进行非极大值抑制处理的操作, 以便去除同一物体的冗余检测。近期, 研究人员开始研究端对端预测的目标检测方法。该类方法直接端对端地为每个物体预测一个检测框。其中, 最具有代表性的方法是 Carion 等人(2020)提出的基于转换器 Transformer 的目标检测方法 DETR (detection transformer), 如图5所示。DETR 利用卷

积神经网络提取特征, 并基于 Transformer 编解码网络直接预测物体的位置以及分类得分。具体地, DETR 预先设定 N 个查询物体特征, 然后将其与编码网络输出的特征共同送入解码网络生成 N 个预测的物体特征, 最后利用预测头网络进行分类和回归。在训练过程中, 为了能够将物体与预测结果一对一匹配, DETR 采用基于匈牙利算法的二分匹配损失。

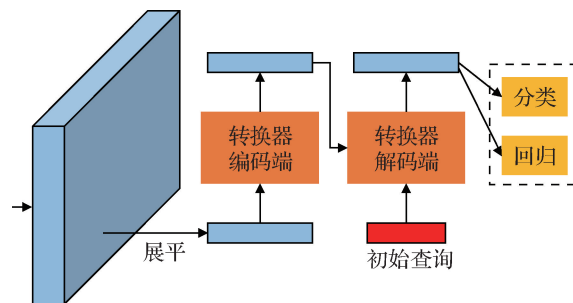


图5 端到端预测的目标检测方法基本架构

Fig. 5 Architecture of end-to-end object detection

尽管 DETR 在目标检测上取得了巨大的成功, 但是它存在收敛速度较慢、小尺度目标检测性能相对较差等问题。为了克服 DETR 存在的问题, Zhu 等人(2021)受形变卷积(Dai 等, 2017)启发提出了 Deformable DETR。与 DETR 采用基于全局的注意力机制不同, Deformable DETR 提出了基于局部稀疏的可形变注意力模块。可形变注意力模块仅关注少量采样位置以及它们之间的关系。同时, 可形变注意力模块可以扩展成多尺度可形变注意力模块, 用在特征金字塔结构中。基于此, Deformable DETR 具有更快的收敛速度、更好的小尺度目标检测性能。类似地, 为了解决 DETR 训练慢、小尺度物体检测性能差等问题, Dai 等人(2021b)提出基于可形变金字塔卷积的编码网络和基于 RoI 池化操作的动态解码网络。受预训练 Transformer 在自然语言中的成功启发, Dai 等人(2021d)提出基于无监督预训练的检测器 UP-DETR (unsupervised DETR)。在预训练过程中, UP-DETR 随机从图像中选取一块区域, 并将其特征添加到 DETR 解码网络的查询特征中预测该区域的位置及重建该特征。实验发现该无监督预训练能够加快 DETR 的收敛速度并提升其检测精度。为了减少 DETR 巨大的计算资源开销, Zheng 等人(2020a)提出了基于自适应聚类 Transformer 的目标检测方法。该方法利用局部敏感哈希对查询特征进

行聚类,并根据聚类结果近似计算注意力特征图,从而减少计算消耗量。为了加快训练速度,Sun 等人(2021d)将 DETR 的思想应用到 FCOS 和 R-CNN 检测器中,通过两者的结合加快了 DETR 的收敛速度。

除此之外,研究人员通过对现有无锚点框目标检测器进行改进,同样实现了基于端到端预测的目标检测。Wang 等人(2021a)基于全卷积网络实现了端到端目标检测器 POTO(prediction-aware one-to-one)。在训练过程中,POTO 采用预测敏感的一对一标签分配原则。与此同时,POTO 提出了 3D 最大值滤波提升空间特征的可区分性,进而更好地压缩同一物体的冗余预测。Sun 等人(2021b)提出了类似的端到端检测器,即根据分类和回归损失一对一分配训练标签。Sun 等人(2021c)对提出可学习候选窗口的端到端检测器 Sparse R-CNN。每个可学习候选框对应应有分类得分、检测框位置以及候选框特征。该方法提取可学习候选框对应的 RoI 特征,并利用候选框特征动态生成用于分类和回归的特征。通过多次迭代,Sparse R-CNN 能够实现端到端地检测物体。

2.4 检测子模块设计

深度目标检测方法存在一些常用的检测子模块,其设计方式对检测性能比较重要。本文简单地介绍一些研究人员主要关注的子模块。

2.4.1 特征金字塔结构设计

特征金字塔结构(feature pyramid network, FPN)是目标检测应对物体尺度变化的重要手段,近年来吸引了研究人员的广泛关注。早期,Lin 等人(2017a)通过引入自上而下的连接,增强特征金字塔结构多个预测层的语义级别。此后,研究人员提出了一系列方法试图进一步增强特征融合质量。Liu 等人(2018b)提出双向融合的特征金字塔结构 PAFPN(path aggregation feature pyramid network),同时进行自上而下和自下而上的双向融合加快不同层的信息交互。Pang 等人(2019)提出将特征金字塔多层融合的共同特征用于增强特征金字塔的每一层。Cao 等人(2020b)和 Li 等人(2019c)提出基于孪生网络的多分支金字塔结构,特征金字塔的每一层都经过数量相同参数共享的卷积层增强每个分支的语义级别。Qiao 等人(2021)提出递归特征金字塔结构,将特征金字塔的输入融合到骨干网络二次特征提取。

上述方法在特征融合时大多采用固定融合的方法,

如特征相加。基于此,研究人员提出了自适应融合的方法。Tan 等人(2020b)、Liu 等人(2019)和 Guo 等人(2020a)分别提出自适应加权的特征融合方法 BiFPN(bi-directional FPN)、ASFF(adaptively spatial feature fusion)和 AugFPN。Wang 等人(2020c)通过可形变 3 维卷积将特征金字塔输入进行融合,生成不同尺度的特征图。Zhang 等人(2020a)将 Transformer 用于特征金字塔不同尺度特征融合。Hu 等人(2021)提出基于注意力机制的特征融合机制,增强上下文信息。Zhao 等人(2021)通过建模层内和层间超像素之间的关系实现自适应特征融合。此外,一些研究人员利用网络搜索的方式试图得到最优的特征金字塔结构,如 Auto-FPN(Xu 等,2019)、NAS-FPN(Ghiasi 等,2019)和 OPA-FPN(one-shot path aggregation FPN)(Liang 等,2021)。Chen 等人(2021b)通过膨胀卷积模块和均匀标签分配策略,能够在不使用金字塔结构的情况下保持相当的精度并提升检测速度。

2.4.2 预测头网络设计

预测头网络主要进行分类和回归两类。在两阶段方法中,Faster R-CNN(Ren 等,2015)主要采用共享全连接的方式进行分类和回归。研究人员认为共享全连接进行分类和回归不是最优方案,并提出一些改进方法。Lu 等人(2019)将候选框区域进行扩增,并采用全卷积网络直接预测物体边界点的位置。Wu 等人(2020b)采用全连接层进行分类、采用全卷积层进行回归。Cao 等人(2020a)采用可区分 RoI 池化层进行分类、采用局部稠密预测进行回归。Song 等人(2020)提出生成两个不同候选框分别进行分类和回归,进而从空间维度上将分类和回归任务解耦。Wang 等人(2020a)提出边缘敏感的边界定位方式取代基于回归的方式。

大部分单阶段方法大多采用两个独立的卷积组(如 4 个卷积层)分别用于分类和回归。一些研究人员认为这种方法缺乏对两个任务的关联。Dai 等人(2021a)提出一种动态注意力头网络统一分类和回归任务的特征生成过程,包含了尺度动态注意力机制、空间动态注意力机制以及任务动态注意力机制。任务动态注意力机制能够自适应地选择不同特征用于不同任务。为了更好地关联分类和回归(定位)两个任务,Li 等人(2021b)提出基于边界框预测分布来估计边界框的不确定性,进而引导分类与检

测质量估计。Feng 等人(2021b)提出任务敏感的预测头网络。分类和回归任务首先共享多层特征,然后利用注意力机制和空间校准机制分别提取适合分类和回归的特征。Chi 等人(2020)提出利用 Transformer 解码结构将不同类型的特征融合,提升现有检测器的检测能力。

2.4.3 标签匹配与损失函数设计

大多数目标检测方法根据交并比准则或者距离准则判定样本(如锚点框)的标签(属于哪个物体)。这些匹配准则虽然简单,但不是最优的。Zhang 等人(2020b)系统地分析了基于锚点框的方法和无锚点框的方法,发现基于锚点框的方法和无锚点框的方法定义正负样本方式是造成它们性能差异的重要原因。基于此,Zhang 等人(2020b)提出了基于自适应训练样本选择策略的检测器 ATSS(adaptive training sample selection)。此后,研究人员进一步提出了许多改进方案。Zhang 等人(2019)提出基于学习匹配的 FreeAnchor 方法,实现锚点框与物体动态匹配。基于分类和回归的预测结果,FreeAnchor 构建基于最大似然估计的损失函数实现动态匹配。Ke 等人(2020)提出多锚点框学习方法 MAL(multiple anchor learning),根据分类和回归的联合分数进行锚点框分配,并提出选择—抑制优化策略防止陷入局部最优。Kim 和 Lee(2020)提出利用混合高斯模型建模物体的概率分布,并根据这个概率进行锚点框的分配。Ma 等人(2021)提出预测样本的检测质量分布,根据检测质量分布进行标签匹配。Ge 等人(2021)将匹配问题转化为最优传输问题,为模糊样本寻找全局最优匹配。端到端预测的目标检测方法(Carion 等,2020)采用匈牙利算法实现一对一最优匹配。Yoo 等人(2021)提出不进行标签匹配,将目标检测转为密度估计问题。

在标签匹配后,检测器基于损失函数进行网络的学习。常用的分类损失函数为交叉熵损失函数和聚焦损失函数(Lin 等,2017b)等。一些研究人员(Chen 等,2019;Qian 等,2020a;Liu 等,2021a;Oksuz 等,2021)将分类问题转换成排序问题,以便更好地解决类别不均衡问题。一些研究人员(Li 等,2021b;Zhang 等,2021a)在设计分类损失函数时考虑其定位精度,以便更好地将分类和回归关联起来。针对回归问题,一些研究人员(Li 等,2020c;Qiu 等,2020a)将回归的偏移量预测转换成边界概率分布

预测问题。除了单独关注分类和回归任务外,一些研究人员(Cao 等,2020c;Wang 和 Zhang,2021b;Gao 等,2021b)研究检测和分类任务的一致性和关联性,其核心思想是希望分类得分高的检测框具有较高的定位精度,以便更好地满足目标检测评测标准。不准确或者错误的标签匹配容易对检测器学习产生不好的影响。针对这一问题,研究人员(Li 等,2019a,2020b;Cai 等,2020)提出一些动态权重调整分类和回归损失的策略减少这些匹配带来的不利影响。

2.4.4 知识蒸馏

知识蒸馏旨在让大网络去引导小网络的学习,帮助小网络在具备较快速度的情况下具有大网络的检测精度。早期,Chen 等人(2017)提出在特征层和预测层进行多层级全特征图逼近。一些研究人员(Li 等,2017;Wang 等,2019c;Sun 等,2020b)认为全特征图逼近容易忽略物体区域,并提出基于物体附近区域的特征逼近方法。研究人员认为背景区域、上下文信息同样有助于辅助提升知识蒸馏的效果。Zhang 和 Ma(2021b)提出注意力引导蒸馏机制和非局部蒸馏机制,缓解前背景信息不平衡和关系利用不充分的问题。Guo 等人(2021a)提出对前景区域和背景区域去耦,分别进行蒸馏。Dai 等人(2021c)提出提取图像中具有可区分力的前景或背景区域进行蒸馏。Chen 等人(2021d)构建候选区域图网络,并对图网络进行蒸馏。Yao 等人(2021)认为两个网络逼近的特征不一定位于金字塔结构的同一层,提出语义引导的自适应特征逼近策略。

3 双目视觉目标检测

单目图像是3维世界中一个视锥的2维投影,丢失了深度信息。双目视觉可以根据物体投影在左右图像上的位置差异计算出视差,并在已知相机参数的情况下根据极线约束计算出像素的深度。在得到每个像素的深度值后,可以逆投影得到视锥中每个像素的3维坐标。因此,双目视觉系统不仅能够预测物体的2维位置和类别信息,还能够预测物体在3维空间中的位置,实现3维目标检测,从而为自动驾驶、工业机器人等任务提供更高层次的场景信息。类似地,利用激光雷达点云检测同样能够实现3维目标检测。与双目视觉目标检测相比,激光雷达点云检测具有更高的检测精度,但是其成本相对

昂贵,对雨雪中等天气变化敏感。本文主要关注双目视觉目标检测方法。受益于深度学习技术的发展,双目目标检测取得了巨大进展。

类似于单目目标检测网络,双目检测网络也可以分为基础骨干、特征融合和预测网络3部分。首先,双目检测通常采用两个权重共享的基础骨干分别得到左右目的单目特征。然后,双目检测进行特征融合,除上文提到的构建特征金字塔外,一般还需要构建双目特征。双目特征构建的方式主要包括直接串接(concatenation)和平面扫描法(plane-sweeping),构建的特征坐标空间属于视锥投影空间。最后,预测网络可以直接使用融合后的视锥空间特征,也可将视锥空间特征显式逆投影到3维空间进行分类和回归。

根据预测网络所使用的特征空间,本文将双目

视觉目标检测方法分为两类:基于直接视锥空间的目标检测方法和基于显式逆投影空间的目标检测方法。基于直接视锥空间的检测过程一般不包含逆投影变换,直接使用基于视锥空间的双目特征进行检测;而基于显式逆投影空间的检测方法一般需要将双目特征进行逆投影变换,生成3维空间上均匀的特征,适合构造体素或转换为俯视图进行检测。图6给出了上述两类方法的发展历程,并给出了一些代表性方法。时间轴上侧为基于直接视锥空间的方法,时间轴下侧为基于显式逆投影空间的方法。图中箭头越长表示该方法在KITTI(Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)数据集(Geiger等,2012)上中等难度(moderate)车辆类别上的平均精度(average precision, AP)越高。

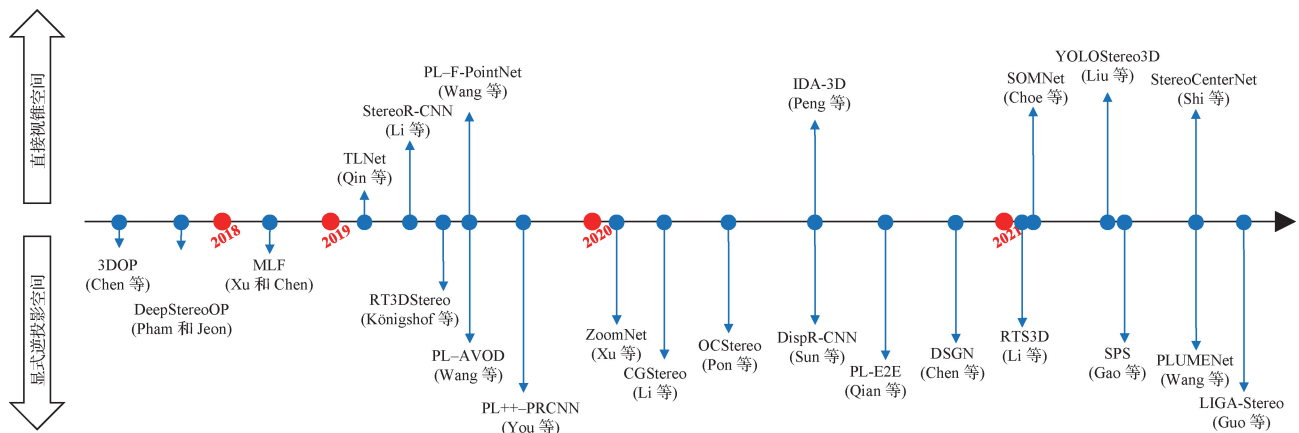


图6 基于深度学习的双目视觉目标检测方法的发展历程

(箭头的长度表示方案在KITTI数据集上的车辆3维检测性能,图中括号内为作者信息)

Fig. 6 Development process of stereo object detection based on deep learning(longer arrow indicates higher Car detection AP on KITTI)

3.1 基于直接视锥空间的双目目标检测方法

基于直接视锥空间的双目目标检测不需要进行额外的坐标空间转换,只需要使用基础骨干提取的两个单目特征构造双目特征。现有方法主要通过串接和平面扫描两种方式构造视锥空间的双目特征。

3.1.1 基于串接构造视锥空间特征的方法

基于串接构造视锥空间特征的方法将基础骨干提取的两个单目视锥空间特征串接起来,利用卷积神经网络强大的拟合能力提取候选框或直接检测3维目标。串接操作不改变原单目特征的坐标空间,是一种简单快速的视锥空间双目特征构造方式。

Li 等人(2019b)提出两阶段方法 Stereo R-CNN。如图7(a),在第1阶段, Stereo R-CNN 利用串接特征得到左右两个成对的候选框。在第2阶

段, Stereo R-CNN 分别提取左右目的 RoI 特征,再次串接特征进行回归。为了得到3维框顶点在左目 RoI 特征内的投影, Stereo R-CNN 引入了一种简化的关键点检测网络,利用得到的关键点信息对最小化投影误差进行数值求解,从而得到质量较高的3维目标检测结果。Shi 等人(2022)借鉴 Stereo R-CNN 双目候选框定义双目包围框,并提出类似于 CenterNet(Zhou 等,2019a)的单阶段无锚点框双目检测方法 StereoCenterNet。StereoCenterNet 在串接的双目特征上预测双目2维框和3维框的朝向、尺寸、底面顶点等信息。预测这些信息后, StereoCenterNet 根据物体遮挡程度不同采用对应的策略来进行最小化投影误差求解,提高了严重遮挡物体的检测精度。

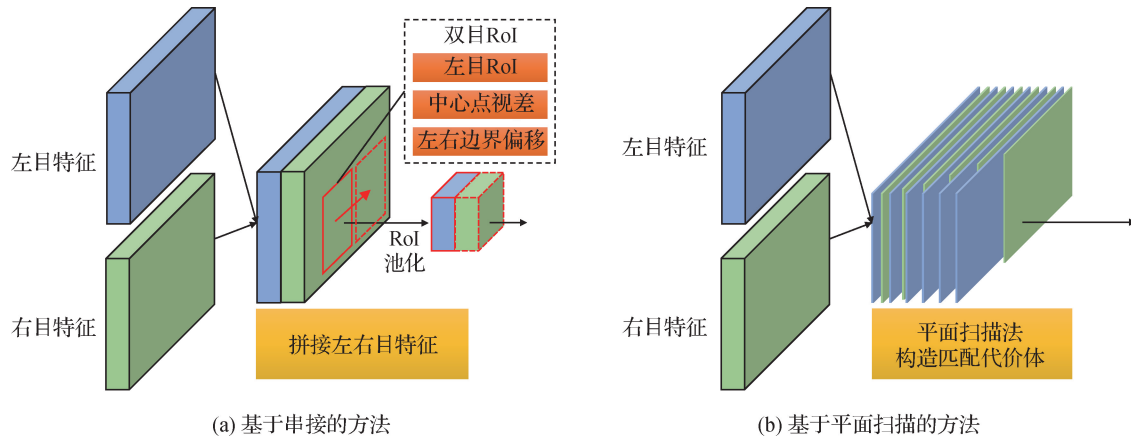


图 7 基于直接视锥空间的双目标检测方法

Fig. 7 Frustum-based stereo 3D detectors ((a) concatenation based method; (b) plane-sweeping based method)

Qin 等人(2019)提出了一种基于 3 维锚点框的两阶段方法。首先,Qin 等人(2019)将 3 维锚点框投影到左右目特征上,得到成对的 2 维候选框。然后,Qin 等人(2019)认为同一目标的左右目 RoI 特征应该相似,提出用余弦相似性对两组特征的每个特征图重新加权,从而增强左右相似的特征图、抑制左右差异较大的特征图。最后,Qin 等人(2019)使用权重调整后的特征进行分类和回归,并在视锥空间对预测结果进行监督,完成 3 维目标检测。

3.1.2 基于平面扫描构造视锥空间特征的方法

为了更好地基于左右目特征提取立体信息,双目深度估计 (Chang 和 Chen, 2018; Xu 和 Zhang, 2020)广泛采用平面扫描法构造匹配代价体 (cost volume)。通过逐视差平面或者逐深度平面地扫描一对 2 维特征,所得 3 维特征即匹配代价体。每一次扫描不改变 2 维特征的坐标空间,所以得到的匹配代价体仍然属于视锥空间。基于平面扫描的检测方法受益于双目深度估计方法的发展,能够直接利用点云监督取得更好的匹配结果,进而学习到每个视锥空间像素是否被物体占据的信息,辅助提高 3 维检测性能。

Peng 等人(2020)提出两阶段的方法 IDA-3D (instance-depth-aware 3D detection),在双目候选框提取阶段使用了与 Stereo R-CNN 相同的提取方式,然后在串接后的左右目 RoI 特征上预测了物体 2 维和 3 维信息,没有再使用额外的关键点检测和投影误差最小化方法。此外,IDA-3D 在左右目融合 RoI 特征上,基于平面扫描法构建了目标级的视差匹配代价体,使用 3 维卷积回归出目标的整体视

差,并以此求得目标整体深度作为 3 维目标的深度坐标。

Liu 等人(2021b)提出一种高效的单阶段方法 YOLOStereo3D。YOLOStereo3D 基于平面扫描法高效地构建了匹配代价体金字塔,使用 Ghost 模块 (Han 等,2020)来快速地增加各级匹配代价体特征的通道数,并在最小分辨率的特征上进行全图视差估计,最后融合同分辨率的左目特征进行 2 维和 3 维目标检测。

Choe 等人(2021)提出 3 维目标检测辅助双目视差估计的方法 SOMNet (stereo object matching network)。其使用的检测方案可视为基于 3 维候选框的两阶段目标检测。在候选框提取阶段,SOMNet 使用平面扫描法构造视锥空间匹配代价体特征,并在其上预测 3 维候选框。在第 2 阶段,SOMNet 提出基于空间占用的候选框特征融合机制,通过估计的视差获得 3 维 RoI 中每个像素是否被物体占用的信息,并构建目标级注意力特征增强视锥空间 RoI,使之更加聚焦于目标表面和形状。

3.2 基于显式逆投影空间的双目标检测方法

在自动驾驶等场景中,感兴趣的目标(如车辆、行人和骑行者等)在 3 维空间中不存在重叠。因此,将存在尺度变化和遮挡问题的视锥空间图像逆投影到尺度均匀、不存在重叠遮挡的 3 维空间,能够缓解视锥投影带来的问题。此外,考虑俯视方向上不存在遮挡问题,还可以把 3 维空间压缩至俯视 2 维空间,在保证性能的同时进一步简化预测网络。逆投影变换主要可以应用在输入图像、特征和候选区域 3 个不同环节。图 8 给出了 3 种方案的示意图。

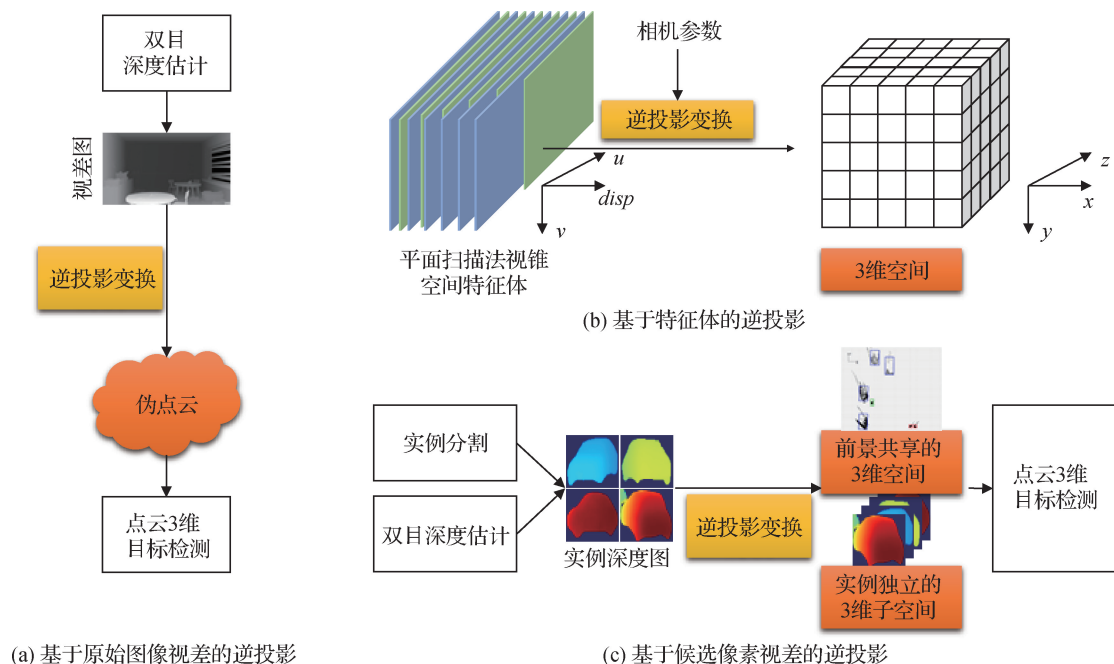


图8 基于显式逆投影空间的双目检测

Fig. 8 Inverse-projection-based stereo 3D detectors ((a) based on inverse-projecting raw disparity maps; (b) based on inverse-projecting feature volumes; (c) based on inverse-projecting instance-level disparity maps)

3.2.1 基于原始图像视差的逆投影方法

基于原始图像视差的逆投影先利用双目视差估计算法预测出逐像素的视差,将这些像素逆投影到3维空间生成点云形式,从而使用相对成熟的点云3维检测方法进行双目目标检测。将这种点云称为伪点云,这种双目目标检测方法称为伪雷达(pseudo-LiDAR, PL)方法。如图8(a)所示,伪雷达方法级联了双目深度估计和点云3维目标检测两个模块,可以使用这两个领域的先进成果协同完成检测。

早期,研究人员先用传统方法完成3维候选框提取,然后用小规模卷积神经网络得到最终的3维检测结果。Chen等人(2015)提出3DOP(3D object proposal),使用已有方法(Yamaguchi等,2014)从双目图像估计出点云,然后求解马尔可夫随机场能量函数最小化问题得到3维候选框,最后利用卷积神经网络对候选框进行优化和评分。在此基础上,Pham和Jeon(2017)提出DeepStereoOP网络,结合图像和深度特征对候选框进行更准确排序。

作者提出MLF(multi-level fusion),是首个使用深度学习技术完成全部预测过程的双目视觉目标检测方法。MLF使用DispNet(Mayer等,2016)从双目图像估计视差图,并逆投影成为深度图和点云,然后

使用2维候选框和点云共同预测物体3维信息。对比了双目和单目输入对3维目标检测的影响,指出双目信息能够明显提高3维检测性能。

Wang等人(2019d)将深度估计和点云3维目标检测方法结合起来,先根据深度生成伪点云,再利用基于点云的检测方法检测3维目标。该方法利用深度估计和点云3维目标检测领域的先进成果,通过级联的方式完成双目3维目标检测任务,性能提升明显。但是,这种级联的方式容易造成误差累积。针对这一问题,You等人(2020)利用真实点云校正伪点云。Li等人(2020a)利用深度估计的置信度来引导深度估计网络,并加上额外的语义分割监督,显著提高了图像中前景深度估计的精度。Peng等人(2022)提出基于两分支网络的方法SIDE(structure-aware instance depth estimation)。一个分支进行基于伪雷达的双目目标检测。另一个分支进行基于视锥空间特征的目标级深度估计,取代点云监督。基于上述两分支结构,SIDE在不需要额外的点云监督的条件下提高了双目3维目标检测性能。上述伪雷达方法需要单独训练深度估计模块和3维目标检测模块,无法进行端到端训练。Qian等人(2020b)通过可微的采样和量化模块设计出端到端训练的伪雷

达方法,称为 PL-E2E(end-to-end PL)。

3.2.2 基于特征体的逆投影方法

上述基于原始图像视差的方法生成伪点云时丢弃了图像提供的颜色和纹理信息,而且没有利用视差估计网络的中间特征。基于特征体的逆投影方法则复用了这些图像特征。具体地,如图8(b)所示,基于特征体逆投影的双目标检测方法通过插值和采样的方式将平面扫描得到的匹配代价体变换到3维空间,利用图像特征提供的颜色和纹理信息,实现了端到端训练的双目标检测。

Chen 等人(2020c)提出一种单阶段的双目3维目标检测方法 DSGN(deep stereo geometry network),使用 PSMNet(pyramid stereo matching network)(Chang 和 Chen,2018)提取匹配代价体并预测深度图,将匹配代价体逆投影得到基于3维空间的几何特征体,并通过3维卷积将其压缩成为俯视图,最后在俯视图上直接进行分类与回归。Guo 等人(2021b)认为,双目视觉中经过特征体逆变换得到的3维空间特征应当与点云检测中的3维空间体素特征相似,同时二者的俯视图特征也应当相似。基于此,Guo 等人(2021b)提出 LIGA-Stereo(LiDAR geometry aware representations for stereo-based 3D detector),设计了一种类似于知识蒸馏的技术,引导双目视觉目标检测的特征逼近性能更好的点云3维目标检测特征。Wang 等人(2021d)提出 PLUME(pseudo LiDAR feature volume),将深度估计和3维目标检测两个任务所使用的特征统一到伪雷达特征空间,从而将原本需要两路神经网络完成的两个任务合并为单路网络,提高了检测速度。

Li 等人(2021a)提出 RTS3D(real-time stereo 3D detection),以双目图像和单目3维检测的粗糙3维框作为输入,构造了一个面向感兴趣区域的特征一致性嵌入空间(feature-consistency embedding space, FCE 空间)进行3维目标检测,并使用迭代的方式优化检测结果。Gao 等人(2021a)改进了 RTS3D 的采样方式,提出基于目标形状先验的非均匀采样获取更多物体表面和周围信息,并设计了一种利用抽象语义信息增强 FCE 空间特征。

3.2.3 基于候选像素视差的逆投影方法

基于原始图像视差的逆投影方法生成了全空间的点云,基于特征体的逆投影方法生成了全空间的3维特征。因此,二者逆投影得到的3维空间包含

了前景目标部分和背景部分。基于候选像素视差的逆投影方法仅聚焦感兴趣目标区域的3维空间(如图8(c)所示),先利用实例分割等方案得到目标的前景像素,然后生成仅含前景区域的3维空间。这种逆投影方法生成的3维空间有效体素较少,可以在有限的检测时间内更灵活地控制特征的空间分辨率;聚焦于前景目标,能够避免不准确的深度估计带来的性能下降。

第1种逆投影策略是去除背景点云、仅保留前景点云。本文称为前景共享的3维空间策略。Königshof 等人(2019)提出一种基于俯视网格图的方法。该方法先使用双目图像预测深度图、2维语义分割和2维包围框得到物体前景部分的深度信息,然后使用深度优先搜索(depth first search,DFS)求解3维空间的连通域,最后在前景共享的3维空间俯视图上回归目标的3维信息。Pon 等人(2020)认为已有的深度估计方法得到的结果在目标边界和形状上并不准确,对3维目标检测有害无益。基于此,Pon 等人(2020)提出面向目标(object-centric, OC)的伪雷达方法,通过增加实例分割模块提取前景共享的3维空间伪点云。

另外一种逆投影策略是为每个实例生成互相独立的3维子空间,即每个3维子空间仅检测单个目标。本文称为实例独立的3维子空间。Xu 等人(2020)提出 ZoomNet,利用2维实例分割和双目深度估计生成基于伪点云的实例独立子空间。图像中,小尺度目标覆盖的像素点比较少,生成点云图比较稀疏,不利于3维检测。为解决该问题,ZoomNet 使用汽车通用3D模型自适应地生成密度一致的子空间点云,提升小尺度目标的检测效果。Sun 等人(2020a)提出 Disp R-CNN,基于候选区域逆投影生成实例独立子空间,并使用自动生成实例3D模型,避免手工标注汽车模型。

4 国内外研究进展比较

视觉目标检测是计算机视觉的经典任务,得到了国内外研究人员的广泛关注。本节简要分析和对比国内外在视觉目标检测方面的研究进展。

4.1 单目视觉目标检测技术

国内外研究机构在单目视觉目标检测方面开展了大量的研究工作。早期国外在基于深度学习的视

觉目标检测技术方面开展了更多更具有代表性的研究工作,如两阶段目标检测器 R-CNN 系列、单阶段目标检测器 YOLO、端到端目标检测器 DETR。近年国内开始在深度目标检测技术方面,特别是单阶段目标检测技术和端到端目标检测技术方面,持续发力,当前已经取得了与国外相当的竞争力。当前国内高水平研究工作(如发表在会议 CVPR (IEEE Conference on Computer Vision and Pattern Recognition) 上)的论文数量方面存在一定的优势,但是在代表性工作方面相对欠缺,存在较大的发展与提升空间。在目标检测开源方面,国内香港中文大学开源的 mmdetection 和国外 Facebook 开源的 detectron2 是目前学术界比较流行的目标检测库。

国内在单目目标检测方面具有代表性的研究机构包括清华大学、中国科学院、香港中文大学、商汤、旷视和华为等。国外在单目目标检测方面具有代表性的研究机构包括斯坦福大学、牛津大学、加州大学伯克利分校、Google 和 Facebook 等。

4.2 双目视觉目标检测技术

清华大学与加拿大多伦多大学合作,在 2015 年率先利用深度学习技术完成 3 维物体检测。2018 年以来,以香港中文大学、中国科学院大学、中国科技大学、浙江大学和武汉大学为代表的国内研究机构,以美国康奈尔大学、德国信息技术研究中心和加拿大多伦多大学为代表的国外研究机构,基于显式逆投影空间设计了伪雷达、DSGN 和 Disp R-CNN 等双目目标检测方法;2019 年以来,香港科技大学、大连理工大学、北京科技大学、美国微软研究院和韩国科学技术院等研究机构提出了多种基于直接视锥空间的双目目标检测方法。

目前,双目视觉目标检测主要应用于无人机和自动驾驶等领域,国内的百度、大疆和美国 Uber、Waabi 等相关企业正在积极与上述科研机构合作,并取得了一定成果。总体上看,国内科研机构在基于深度学习的双目视觉目标检测领域不但取得了先机,且论文数量和检测性能也保持在先进水平,尤其是香港中文大学和香港科技大学分别在两类方法上达到了当前最佳的检测性能。

5 发展趋势与展望

当前基于深度学习的目标检测技术取得了巨大

成功。尽管如此,目标检测技术仍然存在极大发展空间。展望总结目标检测技术的发展趋势如下:

1) 高效率的端到端目标检测。当前基于转换器 Transformer 的端到端目标检测技术取得了一定成功,为目标检测领域的发展注入了新的活力。相比于之前基于锚点框和无锚点框的目标检测方法,该类方法存在收敛减慢、计算资源消耗大等问题。同时,相比于卷积神经网络,Transformer 在计算效率等方面存在一定的劣势。近期,相关研究工作 Deformable DETR (Zhu 等,2021) 和 TSP-FCOS (transformer-based set prediction with FCOS) (Sun 等,2021d) 在一定程度上缓解了这些问题,但是如果设计高效率的 Transformer 编解码网络乃至 Transformer 基础网络进行端到端目标检测仍是未来需要研究的内容之一。

2) 基于自监督学习的目标检测。自监督学习在大规模图像分类任务上取得了与全监督学习相当的性能 (Chen 等,2020b; He 等,2020; Chen 和 He,2021c)。自监督学习用于图像分类任务的前提假设是图像内容被单一物体主导。与图像分类任务不同,目标检测任务中存在数量、尺度不确定的若干物体。因此,如何更好地将自监督学习间接或直接用于目标检测是一个挑战性问题。Xie 等人 (2021) 提出局部和全局的多级监督训练策略提升无监督学习在检测任务上的性能。Liu 等人 (2020a) 利用推土机距离度量不同变换局部位置的相似性,进而进行自监督目标检测。Dai 等人 (2021d) 预测随机子区域在图像的位置并重建随机子区域,实现目标检测的自监督预训练。

3) 长尾分布目标检测。当前目标检测方法大多面向物体检测数据库 MS COCO (Microsoft common objects in context) (<https://cocodataset.org/>) (Lin 等,2014) 和 PASCAL VOC (pattern analysis, statistical modeling and computational learning visual object classes) (<http://www.host.robots.ox.ac.uk/pascal/VOC/>) (Everingham 等,2010)。这两个数据库对于物体的类别有限且不同类别的目标相对均衡充足。然而,现实世界中,物体的类别数量十分庞大且不同类别的物体数量存在极度不平衡,呈现长尾分布现象。针对这一现象,Gupta 等人 (2019) 构建了包含 1 000 类物体的大规模长尾分布数据集 (<https://www.lvisdataset.org/>)。研究人员开始研究相关问

题,并在样本采样(Wu等,2020a)、分组训练(Wu等,2020a;Li等,2020e)、损失函数(Tan等,2020a,2021;Feng等,2021a)等方面开展了相关工作。

4)小样本、零样本目标检测。小样本、零样本目标检测主要关注如何提升训练样本较少甚至没有的物体类别在测试过程中的检测性能。目标检测方法在小样本或零样本情况下的检测能力是通用性的重要标志,是开放世界目标检测(Joseph等,2021)必备的能力。因此,小样本、零样本目标检测具有重要的研究价值。研究人员(Kang等,2019;Zhang等,2021c;Sun等,2021a;Chen等,2021a)在相关方面开展了大量研究,为小样本、零样本目标检测打下坚实的基础。

5)大规模双目目标检测数据集。缺少大规模、高质量双目标注的公开数据集,是当前双目目标检测面临的主要挑战之一。双目目标检测数据集不仅需要标注物体的2维、3维信息,而且需要标注视差、相机参数等。此外,当前很多方法使用了雷达点云、语义分割和实例分割等额外标注信息。因此,建立大规模的双目视觉数据集,并提供高质量的双目标注、完善的评价体系以及开放的测试平台能够为未来双目目标检测发展提供基础性支撑。

6)弱监督双目目标检测。如上所述,建立大规模高质量的双目目标检测数据集是一个复杂且昂贵的系统工程。研究如何在没有高质量双目标注的情况下利用双目数据实现精准3维目标检测十分必要。因此,弱监督的双目目标检测是一个十分重要且具有挑战性的研究方向。

致 谢 本文由中国图象图形学学会视频图像与安全专业委员会组织撰写,该专委会更多详情请见链接:<http://www.csig.org.cn/detail/2448>。

参考文献(References)

- Beal J, Kim E, Tzeng E, Park D H, Zhai A and Kislyuk D. 2020. Toward transformer-based object detection [EB/OL]. [2020-12-17]. <https://arxiv.org/pdf/2012.09958.pdf>
- Bochkovskiy A, Wang C Y and Mark Liao H Y. 2020. Yolov4: optimal speed and accuracy of object detection [EB/OL]. [2020-04-23]. <https://arxiv.org/pdf/2004.10934.pdf>
- Bodla N, Singh B, Chellappa R and Davis L S. 2017. Soft-NMS-Improving object detection with one line of code//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE: 5562-5570 [DOI: 10.1109/ICCV.2017.593]
- Cai Q, Pan Y W, Wang Y, Liu J G, Yao T and Mei T. 2020. Learning a unified sample weighting network for object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 14161-14170 [DOI: 10.1109/CVPR42600.2020.01418]
- Cai Z W and Vasconcelos N. 2018. Cascade R-CNN: Delving into high quality object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 6154-6162 [DOI: 10.1109/CVPR.2018.00644]
- Cao J L, Cholakkal H, Anwer R M, Khan F S, Pang Y W and Shao L. 2020a. D2Det: towards high quality object detection and instance segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 11482-11491 [DOI: 10.1109/CVPR42600.2020.01150]
- Cao J L, Pang Y W, Han J G and Li X L. 2019b. Hierarchical shot detector//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South); IEEE: 9704-9713 [DOI: 10.1109/ICCV.2019.00980]
- Cao J L, Pang Y W and Li X L. 2019a. Triply supervised decoder networks for joint detection and segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 7392-7401 [DOI: 10.1109/CVPR.2019.00757]
- Cao J L, Pang Y W, Zhao S J and Li X L. 2020b. High-level semantic networks for multi-scale object detection. IEEE Transactions on Circuits and Systems for Video Technology, 30 (10): 3372-3386 [DOI: 10.1109/TCSVT.2019.2950526]
- Cao Y H, Chen K, Loy C C and Lin D H. 2020c. Prime sample attention in object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 11580-11588 [DOI: 10.1109/CVPR42600.2020.01160]
- Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK; Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]
- Chang J R and Chen Y S. 2018. Pyramid stereo matching network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 5410-5418 [DOI: 10.1109/CVPR.2018.00567]
- Chen D J, Hsieh H Y and Liu T L. 2021a. Adaptive image transformer for one-shot object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 12247-12256 [DOI: 10.1109/CVPR46437.2021.01207]
- Chen G B, Choi W, Yu X, Han T and Chandraker M. 2017. Learning efficient object detection models with knowledge distillation//Proceedings of the 30st Conference on Neural Information Processing Systems. Long Beach, USA; Curran Associates Inc.: 742-751

- Chen K A, Li J G, Lin W Y, See J, Wang J, Duan L Y, Chen Z B, He C W and Zou J N. 2019. Towards accurate one-stage object detection with AP-loss//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 5119-5127 [DOI: 10.1109/CVPR.2019.00526]
- Chen P G, Liu S, Zhao H S and Jia J Y. 2020a. GridMask data augmentation [EB/OL]. [2022-01-18]. <https://arxiv.org/pdf/2001.04086.pdf>
- Chen Q, Wang Y M, Yang T, Zhang X Y, Cheng J and Sun J. 2021b. You only look one-level feature//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 13039-13048 [DOI: 10.1109/CVPR46437.2021.01284]
- Chen T, Kornblith S, Norouzi M and Hinton G. 2020b. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria; PMLR: 1597-1607
- Chen X L and He K M. 2021c. Exploring simple Siamese representation learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 15750-15758 [DOI: 10.1109/CVPR46437.2021.01549]
- Chen X Z, Kundu K, Zhu Y K, Berneshawi A G, Ma H M, Fidler S and Urtasun R. 2015. 3D object proposals for accurate object class detection//Proceedings of the 28th Conference on Neural Information Processing Systems. Montreal, Canada; Curran Associates Inc.: 424-432 [DOI: 10.5555/2969239.2969287]
- Chen Y H, Zhang Z, Cao Y, Wang L, Lin S and Hu H. 2020e. Rep-Points v2: verification meets regression for object detection//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada; Curran Associates Inc.: 5621-5631
- Chen Y K, Li Y W, Kong T, Qi L, Chu R H, Li L and Jia J Y. 2021e. Scale-aware automatic augmentation for object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 9563-9572 [DOI: 10.1109/CVPR46437.2021.00944]
- Chen Y K, Zhang P Z, Li Z M, Li Y W, Zhang X Y, Qi L, Sun J and Jia J Y. 2020d. Dynamic scale training for object detection [EB/OL]. [2020-03-14]. <https://arxiv.org/pdf/2004.12432.pdf>
- Chen Y L, Liu S, Shen X Y and Jia J Y. 2020c. DSGN: Deep stereo geometry network for 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 12533-12542 [DOI: 10.1109/CVPR42600.2020.01255]
- Chen Y X, Chen P G, Liu S, Wang L W and Jia J Y. 2021d. Deep structured instance graph for distilling object detectors//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada; IEEE: 4359-4368 [DOI: 10.1109/ICCV48922.2021.00432]
- Chi C, Wei F Y and Hu H. 2020. RelationNet ++: bridging visual representations for object detection via transformer decoder//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada; Curran Associates Inc.: 13564-13574
- Choe J, Joo K, Rameau F and So Kweon I S. 2021. Stereo object matching network//Proceedings of 2021 IEEE International Conference on Robotics and Automation. Xi'an, China; IEEE: 12918-12924 [DOI: 10.1109/ICRA48506.2021.9562027]
- Dai J F, Li Y, He K M and Sun J. 2016. R-FCN: object detection via region-based fully convolutional networks//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain; Curran Associates Inc.: 379-387 [DOI: 10.5555/3157096.3157139]
- Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE: 764-773 [DOI: 10.1109/ICCV.2017.89]
- Dai X, Jiang Z R, Wu Z, Bao Y P, Wang Z C, Liu S and Zhou E J. 2021c. General instance distillation for object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 7842-7851 [DOI: 10.1109/CVPR46437.2021.00775]
- Dai X Y, Chen Y P, Xiao B, Chen D D, Liu M C, Yuan L and Zhang L. 2021a. Dynamic head: unifying object detection heads with attentions//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 7373-7382 [DOI: 10.1109/CVPR46437.2021.00729]
- Dai X Y, Chen Y P, Yang J W, Zhang P C, Yuan L and Zhang L. 2021b. Dynamic DETR: End-to-end object detection with dynamic attention//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada; IEEE: 2988-2997 [DOI: 10.1109/ICCV48922.2021.00298]
- Dai Z G, Cai B L, Lin Y G and Chen J Y. 2021d. UP-DETR: unsupervised pre-training for object detection with transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 1601-1610 [DOI: 10.1109/CVPR46437.2021.00165]
- Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA; IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]
- DeVries T and Taylor G W. 2017. Improved regularization of convolutional neural networks with cutout [EB/OL]. [2022-01-18]. <https://arxiv.org/pdf/1708.04552.pdf>
- Dong Z W, Li G X, Liao Y, Wang F, Ren P J and Qian C. 2020. CentripetalNet: pursuing high-quality Keypoint pairs for object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 10516-10525 [DOI: 10.1109/CVPR42600.2020.01053]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houtsby N. 2021. An image is worth 16 x 16 words;

- Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: <https://openreview.net/forum?id=YicbFdNTTy>
- Duan K W, Bai S, Xie L X, Qi H G, Huang Q M and Tian Q. 2019. CenterNet: keypoint triplets for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE: 6568-6577 [DOI: 10. 1109/ICCV. 2019. 00667]
- Duan K W, Xie L X, Qi H G, Bai S, Huang Q M and Tian Q. 2020. Corner proposal network for anchor-free, two-stage object detection//Proceedings of 2020 European Conference on Computer Vision. Glasgow, UK: Springer: 399-416 [DOI: 10. 1007/978-3-030-58580-8_24]
- Dvornik N, Shmelkov K, Mairal J and Schmid C. 2017. BlitzNet: A real-time deep network for scene understanding//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4174-4182 [DOI: 10. 1109/ICCV. 2017. 447]
- Everingham M, Van Gool L, Williams C K I, Winn J and Zisserman A. 2010. The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision, 88(2): 303-338 [DOI: 10. 1007/s11263-009-0275-4]
- Fang H S, Sun J H, Wang R Z, Gou M H, Li Y L and Lu C W. 2019. InstaBoost: boosting instance segmentation via probability map guided copy-pasting//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 682-691 [DOI: 10. 1109/ICCV. 2019. 00077]
- Feng C J, Zhong Y J, Gao Y, Scott M R and Huang W L. 2021b. TOOD: task-aligned one-stage object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3510-3519 [DOI: 10. 1109/ICCV48922. 2021. 00349]
- Feng C J, Zhong Y J and Huang W L. 2021a. Exploring classification equilibrium in long-tailed object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3417-3426 [DOI: 10. 1109/ICCV48922. 2021. 00340]
- Fu C Y, Liu W, Ranga A, Tyagi A and Berg A C. 2017. DSSD: deconvolutional single shot detector [EB/OL]. [2022-01-18]. <https://arxiv.org/pdf/1701.06659.pdf>
- Gao A Q, Cao J L and Pang Y W. 2021a. Shape prior non-uniform sampling guided real-time stereo 3D object detection [EB/OL]. [2021-06-22]. <https://arxiv.org/pdf/2106.10013.pdf>
- Gao Z T, Wang L M and Wu G S. 2021b. Mutual supervision for dense object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3641-3650 [DOI: 10. 1109/ICCV48922. 2021. 00362]
- Ge Z, Liu S T, Li Z M, Yoshie O and Sun J. 2021. OTA: optimal transport assignment for object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 303-312 [DOI: 10. 1109/CVPR46437. 2021. 00037]
- Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 3354-3361 [DOI: 10. 1109/CVPR. 2012. 6248074]
- Ghiasi G, Lin T Y and Le Q V. 2019. NAS-FPN: Learning scalable feature pyramid architecture for object detection//Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 7036-7045 [DOI: 10. 1109/CVPR. 2019. 00720]
- Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10. 1109/ICCV. 2015. 169]
- Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10. 1109/CVPR. 2014. 81]
- Guo C X, Fan B, Zhang Q, Xiang S M and Pan C H. 2020a. AugFPN: Improving multi-scale feature learning for object detection//Proceedings of 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12592-12601 [DOI: 10. 1109/CVPR42600. 2020. 01261]
- Guo J Y, Han K, Wang Y H, Wu H, Chen X H, Xu C J and Xu C. 2021a. Distilling object detectors via decoupled features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2154-2164 [DOI: 10. 1109/CVPR46437. 2021. 00219]
- Guo J Y, Han K, Wang Y H, Zhang C, Yang Z H, Wu H, Chen X H and Xu C. 2020b. Hit-detector: hierarchical trinity architecture search for object detection//Proceedings of 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11402-11411 [DOI: 10. 1109/CVPR42600. 2020. 01142]
- Guo X Y, Shi S S, Wang X G and Li H S. 2021b. LIGA-stereo: learning Lidar geometry aware representations for stereo-based 3D detector//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3153-3163
- Gupta A, Dollar P and Girshick R. 2019. LVIS: a dataset for large vocabulary instance segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5356-5364 [DOI: 10. 1109/CVPR. 2019. 00550]
- Han K, Wang Y H, Tian Q, Guo J Y, Xu C J and Xu C. 2020. GhostNet: more features from cheap operations//Proceedings of 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1577-1586 [DOI: 10. 1109/CVPR42600. 2020. 00165]
- He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and

- Pattern Recognition. Seattle, USA; IEEE: 9726-9735 [DOI: 10.1109/CVPR42600.2020.00975]
- He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]
- He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- He Y H, Zhang X Y, Savvides M and Kitani K. 2018. Softer-NMS: rethinking bounding box regression for accurate object detection [EB/OL]. [2022-01-18]. <http://arxiv.org/pdf/1809.08545.pdf>
- Hu M, Li Y L, Fang L and Wang S J. 2021. A²-FPN: attention aggregation based feature pyramid network for instance segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 15343-15352 [DOI: 10.1109/CVPR46437.2021.01509]
- Huang G, Liu Z, van der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]
- Huang L C, Yang Y, Deng Y F and Yu Y N. 2015. DenseBox: unifying landmark localization with end to end object detection [EB/OL]. [2022-01-18]. <http://arxiv.org/pdf/1509.04874.pdf>
- Jiang B R, Luo R X, Mao J Y, Xiao T T and Jiang Y N. 2018. Acquisition of localization confidence for accurate object detection//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany; Springer: 816-832 [DOI: 10.1007/978-3-030-01264-9_48]
- Jiang W T, Zhang C, Zhang S C and Liu W J. 2019. Multiscale feature map fusion algorithm for target detection. Journal of Image and Graphics, 24(11): 1918-1931 (姜文涛, 张驰, 张晟翀, 刘万军. 2019. 多尺度特征图融合的目标检测. 中国图象图形学报, 24(11): 1918-1931) [DOI: 10.11834/jig.190021]
- Joseph K J, Khan S, Khan F S and Balasubramanian V N. 2021. Towards open world object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 5830-5840 [DOI: 10.1109/CVPR46437.2021.00577]
- Kang B Y, Liu Z, Wang X, Yu F, Feng J S and Darrell T. 2019. Few-shot object detection via feature reweighting//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South); IEEE: 8419-8428 [DOI: 10.1109/ICCV.2019.00851]
- Ke W, Zhang T L, Huang Z Y, Ye Q X, Liu J Z and Huang D. 2020. Multiple anchor learning for visual object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 10203-10212 [DOI: 10.1109/CVPR42600.2020.01022]
- Kim K and Lee H S. 2020. Probabilistic anchor assignment with IoU prediction for object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK; Springer: 355-371 [DOI: 10.1007/978-3-030-58595-2_22]
- Kim S W, Kook H K, Sun J Y, Kang M C and Ko S J. 2018. Parallel feature pyramid network for object detection//Proceedings of the 15th Conference on Computer Vision. Munich, Germany; Springer: 239-256 [DOI: 10.1007/978-3-030-01228-1_15]
- Kong T, Sun F C, Huang W B and Liu H P. 2018. Deep feature pyramid reconfiguration for object detection//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany; Springer: 172-188 [DOI: 10.1007/978-3-030-01228-1_11]
- Kong T, Sun F C, Liu H P, Jiang Y N, Li L and Shi J B. 2019. FoveaBox: Beyond anchor-based object detector [EB/OL]. [2022-01-18]. <http://arxiv.org/pdf/1904.03797.pdf>
- Königshof H, Salscheider N O and Stiller C. 2019. Realtime 3D object detection for automated driving using stereo vision and semantic information//Proceedings of 2019 IEEE Intelligent Transportation Systems Conference. Auckland, New Zealand; IEEE: 1405-1410 [DOI: 10.1109/ITSC.2019.8917330]
- Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th Conference on Neural Information Processing Systems. Lake Tahoe, USA; Curran Associates Inc.: 1106-1114
- Lan S Y, Ren Z, Wu Y, Davis L S and Hua G. 2020. SaccadeNet: A fast and accurate object detector//Proceedings of 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 10394-10403 [DOI: 10.1109/CVPR42600.2020.01041]
- Law H and Deng J. 2018. CornerNet: detecting objects as paired keypoints//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany; Springer: 765-781 [DOI: 10.1007/978-3-030-01264-9_45]
- Law H, Teng Y, Russakovsky O and Deng J. 2020. CornerNet-lite: efficient keypoint based object detection//Proceedings of the British Machine Vision Conference. [s.l.]; BMVA
- Li B Y, Liu Y and Wang X G. 2019a. Gradient harmonized single-stage detector//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA; AAAI: 8577-8584 [DOI: 10.1609/aaai.v33i01.33018577]
- Li C Y, Ku J and Waslander S L. 2020a. Confidence guided stereo 3D object detection with split depth estimation//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, USA; IEEE: 5776-5783 [DOI: 10.1109/IROS45743.2020.9341188]
- Li H D, Wu Z X, Zhu C, Xiong C M, Socher R and Davis L S. 2020b. Learning from noisy anchors for one-stage object detection//Proceed-

- ings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 10585-10594 [DOI: 10.1109/CVPR42600.2020.01060]
- Li H H, Zhou K P, Han T C. 2000. Ship object detection based on SSD improved with CReLU and FPN. Chinese Journal of Scientific Instrument, 41(4): 183-190 (李晖晖, 周康鹏, 韩太初. 2000. 基于CReLU和FPN改进的SSD舰船目标检测. 仪器仪表学报, 41(4): 183-190) [DOI: 10.19650/j.cnki.cjsi.J2006122]
- Li P L, Chen X Z and Shen S J. 2019b. Stereo R-CNN based 3D object detection for autonomous driving//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 7644-7652 [DOI: 10.1109/CVPR.2019.00783]
- Li P X, Su S and Zhao H C. 2021a. RTS3D: real-time stereo 3D detection from 4D feature-consistency embedding space for autonomous driving//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s. l.]: AAAI: 1930-1939
- Li Q Q, Jin S Y and Yan J J. 2017. Mimicking very efficient network for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 7341-7349 [DOI: 10.1109/CVPR.2017.776]
- Li X, Wang W H, Hu X L, Li J, Tang J H and Yang J. 2021b. Generalized focal loss V2: learning reliable localization quality estimation for dense object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 11632-11641 [DOI: 10.1109/CVPR46437.2021.01146]
- Li X, Wang W H, Wu L J, Chen S, Hu X L, Li J, Tang J H and Yang J. 2020c. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada; Curran Associates Inc.: 21002-21012
- Li Y, Wang T, Kang B Y, Tang S, Wang C F, Li J T and Feng J S. 2020e. Overcoming classifier imbalance for long-tail object detection with balanced group Softmax//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 10988-10997 [DOI: 10.1109/CVPR42600.2020.01100]
- Li Y H, Chen Y T, Wang N Y and Zhang Z X. 2019c. Scale-aware trident networks for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South); IEEE: 6053-6062 [DOI: 10.1109/ICCV.2019.00615]
- Li Y Z, Pang Y W, Shen J B, Cao J L and Shao L. 2020d. NETNet: neighbor erasing and transferring network for better single shot object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 13349-13358 [DOI: 10.1109/CVPR42600.2020.01336]
- Liang T T, Wang Y T, Tang Z, Hu G S and Ling H B. 2021. OPANAS: one-shot path aggregation network architecture search for object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 10195-10203 [DOI: 10.1109/CVPR46437.2021.01006]
- Lin T Y, Dollár P, Girshick R, He K M, Hariharan B and Belongie S. 2017a. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 936-944 [DOI: 10.1109/CVPR.2017.106]
- Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017b. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]
- Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland; Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]
- Liu J, Li D, Zheng R Z, Tian L and Shan Y. 2021a. RankDetNet: Delving into ranking constraints for object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 264-273 [DOI: 10.1109/CVPR46437.2021.00033]
- Liu S, Qi L, Qin H F, Shi J P and Jia J Y. 2018b. Path aggregation network for instance segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 8759-8768 [DOI: 10.1109/CVPR.2018.00913]
- Liu S T, Huang D and Wang Y H. 2018a. Receptive field block net for accurate and fast object detection//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany; Springer: 404-419 [DOI: 10.1007/978-3-030-01252-6_24]
- Liu S T, Huang D and Wang Y H. 2019. Learning spatial fusion for single-shot object detection [EB/OL]. [2022-01-18]. <http://arxiv.org/pdf/1911.09516.pdf>
- Liu S T, Li Z M and Sun J. 2020a. Self-EMD: Self-supervised object detection without ImageNet [EB/OL]. [2022-01-18]. <http://arxiv.org/pdf/2011.13677.pdf>
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot multiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands; Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]
- Liu Y X, Wang L J and Liu M. 2021b. YOLOStereo3D: a step back to 2 d for efficient stereo 3D detection//Proceedings of 2021 IEEE/CVF International Conference on Robotics and Automation. Xi'an, China; IEEE: 13018-13024 [DOI: 10.1109/ICRA48506.2021.9561423]
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021c. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada; IEEE: 10012-10022 [DOI: 10.1109/ICCV48922.2021.00986]
- Liu Z L, Zheng T, Xu G D, Yang Z, Liu H F and Cai D. 2020b. Train-

- ing-time-friendly network for real-time object detection//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI; 11685-11692 [DOI: 10.1609/aaai.v34i07.6838]
- Lu X, Li B Y, Yue Y X, Li Q Q and Yan J J. 2019. Grid R-CNN//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE; 7363-7372 [DOI: 10.1109/CVPR.2019.00754]
- Ma Y C, Liu S T, Li Z M and Sun J. 2021. IQDet: instance-wise quality distribution sampling for object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE; 1717-1725 [DOI: 10.1109/CVPR46437.2021.00176]
- Mayer N, Ilg E, Häusser P, Fischer P, Cremers D, Dosovitskiy A and Brox T. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE; 4040-4048 [DOI: 10.1109/CVPR.2016.438]
- Newell A, Yang K Y and Deng J. 2016. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer; 483-499 [DOI: 10.1007/978-3-319-46484-8_29]
- Nie J, Anwer R M, Cholakkal H, Khan F S, Pang Y W and Shao L. 2019. Enriched feature guided refinement network for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE; 9536-9545 [DOI: 10.1109/ICCV.2019.00963]
- Oksuz K, Cam B C, Akbas E and Kalkan S. 2021. Rank and sort loss for object detection and instance segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE; 3009-3018 [DOI: 10.1109/ICCV48922.2021.00300]
- Pang J M, Chen K, Shi J P, Feng H J, Ouyang W L and Lin D H. 2019. Libra R-CNN: towards balanced learning for object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE; 821-830 [DOI: 10.1109/CVPR.2019.00091]
- Pato L V, Negrinho R and Aguiar P M Q. 2020. Seeing without looking: Contextual rescoring of object detections for AP maximization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE; 14598-14606 [DOI: 10.1109/CVPR42600.2020.01462]
- Peng W L, Pan H, Liu H and Sun Y. 2020. IDA-3D: instance-depth-aware 3D object detection from stereo vision for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE; 13012-13021 [DOI: 10.1109/CVPR42600.2020.01303]
- Peng X D, Zhu X G, Wang T and Ma Y X. 2022. SIDE: center-based stereo 3D detector with structure-aware instance depth estimation//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE; 119-128 [DOI: 10.1109/WACV51458.2022.00030]
- Pham C C and Jeon J W. 2017. Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks. *Signal Processing: Image Communication*, 53: 110-122 [DOI: 10.1016/j.image.2017.02.007]
- Pon A D, Ku J, Li C Y and Waslander S L. 2020. Object-centric stereo matching for 3D object detection//Proceedings of 2020 IEEE/CVF International Conference on Robotics and Automation. Paris, France: IEEE; 8383-8389 [DOI: 10.1109/ICRA40945.2020.9196660]
- Qian Q, Chen L, Li H and Jin R. 2020a. DR loss: improving object detection by distributional ranking//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE; 12161-12169 [DOI: 10.1109/CVPR42600.2020.01218]
- Qian R, Garg D, Wang Y, You Y R, Belongie S, Hariharan B, Campbell M, Weinberger K Q and Chao W L. 2020b. End-to-end pseudo-lidar for image-based 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE; 5880-5889 [DOI: 10.1109/CVPR42600.2020.00592]
- Qiao S Y, Chen L C and Yuille A. 2021. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE; 10213-10224 [DOI: 10.1109/CVPR46437.2021.01008]
- Qin Z Y, Wang J L and Lu Y. 2019. Triangulation learning network: from monocular to stereo 3D object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE; 7615-7623 [DOI: 10.1109/CVPR.2019.00780]
- Qiu H, Ma Y C, Li Z M, Liu S T and Sun J. 2020b. BorderDet: Border feature for dense object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer; 549-564 [DOI: 10.1007/978-3-030-58452-8_32]
- Qiu H Q, Li H L, Wu Q B and Shi H C. 2020a. Offset bin classification network for accurate object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE; 13185-13194 [DOI: 10.1109/CVPR42600.2020.01320]
- Qiu H Q, Li H L, Wu Q B, Cui J H, Song Z C, Wang L X and Zhang M J. 2021. CrossDet: crossline representation for object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE; 3195-3204 [DOI: 10.1109/ICCV48922.2021.00318]
- Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: Unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE; 779-788 [DOI: 10.1109/CVPR.2016.91]

- Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA; IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]
- Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2022-01-18]. <https://arxiv.org/pdf/1804.02767.pdf>
- Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks// Proceedings of the 28th Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 91-99
- Rezatofighi H, Tsoi N, Gwak J Y, Sadeghian A, Reid I and Savarese S. 2019. Generalized intersection over union: a metric and a loss for bounding box regression//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 658-666 [DOI: 10.1109/CVPR.2019.00075]
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115 (3): 211-252 [DOI: 10.1007/s11263-015-0816-y]
- Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R and LeCun Y. 2014. OverFeat: Integrated recognition, localization and detection using convolutional networks//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: [s. n.]
- Shi Y G, Guo Y, Mi Z Q and Li X J. 2022. Stereo centerNet-based 3D object detection for autonomous driving. Neurocomputing, 471: 219-229 [DOI: 10.1016/j.neucom.2021.11.048]
- Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2022-01-18]. <https://arxiv.org/pdf/1409.1556.pdf>
- Singh B and Davis L S. 2018. An analysis of scale invariance in object detection-SNIP//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 3578-3587 [DOI: 10.1109/CVPR.2018.00377]
- Singh B, Najibi M and Davis L S. 2018. SNIPER: efficient multi-scale training//Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 9333-9343
- Singh K K and Lee Y J. 2017. Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy; IEEE: 3544-3553 [DOI: 10.1109/ICCV.2017.381]
- Song G L, Liu Y and Wang X G. 2020. Revisiting the sibling head in object detector//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 11560-11569 [DOI: 10.1109/CVPR42600.2020.01158]
- Sun B, Li B H, Cai S C, Yuan Y and Zhang C. 2021a. FSCE: few-shot object detection via contrastive proposal encoding//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [s. l.]; IEEE: 7352-7362
- Sun J M, Chen L H, Xie Y M, Zhang S Y, Jiang Q H, Zhou X W and Bao H J. 2020a. DISP R-CNN: stereo 3D object detection via shape prior guided instance disparity estimation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 10545-10554 [DOI: 10.1109/CVPR42600.2020.01056]
- Sun P Z, Jiang Y, Xie E Z, Shao W Q, Yuan Z H, Wang C H and Luo P. 2021b. What makes for end-to-end object detection//Proceedings of the 38th International Conference on Machine Learning. [s. l.]; PMLR: 9934-9944
- Sun P Z, Zhang R F, Jiang Y, Kong T, Xu C F, Zhan W, Tomizuka M, Li L, Yuan Z H, Wang C H and Luo P. 2021c. Sparse R-CNN: end-to-end object detection with learnable proposals//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 14454-14463 [DOI: 10.1109/CVPR46437.2021.01422]
- Sun R Y, Tang F H, Zhang X P, Xiong H K and Tian Q. 2020b. Distilling object detectors with task adaptive regularization [EB/OL]. [2022-01-18]. <https://arxiv.org/pdf/2006.13108.pdf>
- Sun Z Q, Cao S C, Yang Y M and Kitani K. 2021 d. Rethinking transformer-based set prediction for object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada; IEEE: 3611-3620 [DOI: 10.1109/ICCV48922.2021.00359]
- Szegedy C, Toshev A and Erhan D. 2013. Deep neural networks for object detection//Proceedings of the 26th Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 2553-2561
- Tan J R, Lu X, Zhang G, Yin C Q and Li Q Q. 2021. Equalization loss v2: a new gradient balance approach for long-tailed object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 1685-1694 [DOI: 10.1109/CVPR46437.2021.00173]
- Tan J R, Wang C B, Li B Y, Li Q Q, Ouyang W L, Yin C Q and Yan J J. 2020a. Equalization loss for long-tailed object recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 11659-11668 [DOI: 10.1109/CVPR42600.2020.01168]
- Tan M X, Pang R M and Le Q V. 2020b. EfficientDet: Scalable and efficient object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 10778-10787 [DOI: 10.1109/CVPR42600.2020.01079]
- Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South); IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]
- Uijlings J R R, Van De Sande K E A, Gevers T and Smeulders A W M. 2013. Selective search for object recognition. International Journal

- of Computer Vision, 104 (2): 154-171 [DOI: 10.1007/s11263-013-0620-5]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]
- Viola P and Jones M J. 2004. Robust real-time face detection. International Journal of Computer Vision, 57 (2): 137-154 [DOI: 10.1023/B:VISI.0000013087.49260.fb]
- Vu T, Jang H, Pham T X and Yoo C D. 2019. Cascade RPN: delving into high-quality region proposal network with adaptive convolution//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 1430-1440
- Wang J F, Song L, Li Z M, Sun H B, Sun J and Zheng N N. 2021a. End-to-end object detection with fully convolutional network//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15849-15858 [DOI: 10.1109/CVPR46437.2021.01559]
- Wang J Q, Chen K, Yang S, Loy C C and Lin D H. 2019a. Region proposal by guided anchoring//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2965-2974 [DOI: 10.1109/CVPR.2019.00308]
- Wang J Q, Zhang W W, Cao Y H, Chen K, Pang J M, Gong T, Shi J P, Loy C C and Lin D H. 2020a. Side-aware boundary localization for more precise object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 403-419 [DOI: 10.1007/978-3-030-58548-8_24]
- Wang K Y and Zhang L. 2021b. Reconcile prediction consistency for balanced object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 3631-3640
- Wang N, Gao Y, Chen H, Wang P, Tian Z, Shen C H and Zhang Y N. 2020b. NAS-FCOS: fast neural architecture search for object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11940-11948 [DOI: 10.1109/CVPR42600.2020.01196]
- Wang T, Yuan L, Zhang X P and Feng J S. 2019c. Distilling object detectors with fine-grained feature imitation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4933-4942 [DOI: 10.1109/CVPR.2019.00507]
- Wang T C, Anwer R M, Cholakkal H, Khan F S, Pang Y W and Shao L. 2019b. Learning rich features at high-speed for single-shot object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1971-1980 [DOI: 10.1109/ICCV.2019.00206]
- Wang W H, Xie E Z, Li X, Fan D P, Song K T, Liang D, Lu T, Luo P and Shao L. 2021c. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 568-578 [DOI: 10.1109/ICCV48922.2021.00061]
- Wang X J, Zhang S L, Yu Z R, Feng L T and Zhang W. 2020c. Scale-equalizing pyramid convolution for object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13356-13365 [DOI: 10.1109/CVPR42600.2020.01337]
- Wang Y, Chao W L, Garg D, Hariharan B, Campbell M and Weinberger K Q. 2019d. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8445-8453 [DOI: 10.1109/CVPR.2019.00864]
- Wang Y, Yang B, Hu R, Liang M and Urtasun R. 2021d. PLUMENet: efficient 3D object detection from stereo images//Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, Czech Republic: IEEE: 3383-3390 [DOI: 10.1109/IROS51168.2021.9635875]
- Wei F Y, Sun X, Li H Y, Wang J D and Lin S. 2020. Point-set anchors for object detection, instance segmentation and pose estimation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 527-544 [DOI: 10.1007/978-3-030-58607-2_31]
- Wu J L, Song L C, Wang T C, Zhang Q and Yuan J S. 2020a. Forest R-CNN: large-vocabulary long-tailed object detection and instance segmentation//Proceedings of 2020 ACM Multimedia. Seattle, USA: ACM: 1570-1578 [DOI: 10.1145/3394171.3413970]
- Wu Y, Chen Y P, Yuan L, Liu Z C, Wang L J, Li H Z and Fu Y. 2020b. Rethinking classification and localization for object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10183-10192 [DOI: 10.1109/CVPR42600.2020.01020]
- Xie E Z, Ding J, Wang W H, Zhan X H, Xu H, Sun P Z, Li Z G and Luo P. 2021. DetCo: unsupervised contrastive learning for object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 8392-8401 [DOI: 10.1109/ICCV48922.2021.00828]
- Xu B and Chen Z Z. 2018. Multi-level fusion based 3D object detection from monocular images//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 2345-2353 [DOI: 10.1109/CVPR.2018.00249]
- Xu H, Yao L W, Zhang W, Liang X D and Li Z G. 2019. Auto-FPN: automatic network architecture adaptation for object detection beyond classification//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6648-6657 [DOI: 10.1109/ICCV.2019.00675]
- Xu H F and Zhang J Y. 2020. AANet: adaptive aggregation network for efficient stereo matching//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA:

- 1956-1965 [DOI: 10.1109/CVPR42600.2020.00203]
- Xu Z B, Zhang W, Ye X Q, Tan X, Yang W, Wen S L, Ding E R, Meng A J and Huang L S. 2020. ZoomNet: part-aware adaptive zooming neural network for 3D object detection//Proceedings of 2020 AAAI Conference on Artificial Intelligence. New York, USA; AAAI: 12557-12564
- Yamaguchi K, McAllester D and Urtasun R. 2014. Efficient joint segmentation, occlusion labeling, stereo and flow estimation//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland; Springer: 756-771 [DOI: 10.1007/978-3-319-10602-1_49]
- Yang Z, Liu S H, Hu H, Wang L W and Lin S. 2019. RepPoints: point set representation for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South); IEEE: 9656-9665 [DOI: 10.1109/ICCV.2019.00975]
- Yang Z, Xu Y H, Xue H, Zhang Z, Urtasun R, Wang L W, Lin S and Hu H. 2020. Dense RepPoints: representing visual objects with dense point sets//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK; Springer: 227-244 [DOI: 10.1007/978-3-030-58589-1_14]
- Yao L W, Pi R J, Xu H, Zhang W, Li Z G and Zhang T. 2021. G-DetKD: towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada; IEEE: 3591-3600 [DOI: 10.1109/ICCV48922.2021.00357]
- Yao Q L, Hu X and Lei H. 2019. Aircraft detection in remote sensing imagery with multi-scale feature fusion convolutional neural networks. *Acta Geodaetica et Cartographica Sinica*, 48(10): 1266-1274 (姚群力, 胡显, 雷宏. 2019. 基于多尺度融合特征卷积神经网络的遥感图像飞机目标检测. *测绘学报*, 48(10): 1266-1274) [DOI: 10.11947/j. AGCS.2019.20180398]
- Yoo J, Lee H, Chung I, Seo G and Kwak N. 2021. Training multi-object detector by estimating bounding box distribution for input image//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada; IEEE: 3437-3446 [DOI: 10.1109/ICCV48922.2021.00342]
- You Y R, Wang Y, Chao W L, Garg D, Pleiss G, Hariharan B, Campbell M and Weinberger K Q. 2020. Pseudo-LiDAR ++: accurate depth for 3D object detection in autonomous driving//Proceedings of 2020 International Conference on Learning Representations. Addis Ababa, Ethiopia; OpenReview.net
- Yu J H, Jiang Y N, Wang Z Y, Cao Z M and Huang T. 2016. UnitBox: an advanced object detection network//Proceedings of 2016 ACM Multimedia. Amsterdam, the Netherlands; ACM: 516-520 [DOI: 10.1145/2964284.2967274]
- Yun S, Han D, Chun S, Oh S J, Yoo Y and Choe J. 2019. CutMix: regularization strategy to train strong classifiers with localizable features//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South); IEEE: 6022-6031 [DOI: 10.1109/ICCV.2019.00612]
- Zhang D, Zhang H W, Tang J H, Wang M, Hua X S and Sun Q R. 2020a. Feature pyramid transformer//Proceedings of the 16th European Conference on Computer Vision, 2020. Glasgow, UK; Springer: 323-339 [DOI: 10.1007/978-3-030-58604-1_20]
- Zhang H Y, Cisse M, Dauphin Y N and Lopez-Paz D. 2018a. Mixup: beyond empirical risk minimization//Proceedings of 2018 International Conference on Learning Representations. Vancouver, Canada; OpenReview.net
- Zhang H Y, Wang Y, Dayoub F and Sünderrhauf N. 2021a. VarifocalNet: an IoU-aware dense object detector//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 8514-8523 [DOI: 10.1109/CVPR46437.2021.00841]
- Zhang L, Zhou S G, Guan J H and Zhang J. 2021c. Accurate few-shot object detection with support-query mutual guidance and hybrid loss//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 14424-14432 [DOI: 10.1109/CVPR46437.2021.01419]
- Zhang L F and Ma K S. 2021b. Improve object detection with feature-based knowledge distillation: towards accurate and efficient detectors//Proceedings of the 9th International Conference on Learning Representations. [s. l.]; OpenReview.net
- Zhang S F, Chi C, Yao Y Q, Lei Z and Li S Z. 2020b. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 9756-9765 [DOI: 10.1109/CVPR42600.2020.00978]
- Zhang S F, Wen L Y, Bian X, Lei Z and Li S Z. 2018b. Single-shot refinement neural network for object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 4203-4212 [DOI: 10.1109/CVPR.2018.00442]
- Zhang W, Zhuang X T, Wang X L, Chen Y F and Li Y C. 2021. DS-YOLO: a real-time small object detection algorithm on UAVs. *Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)*, 41(1): 86-98 (张伟, 庄幸涛, 王雪力, 陈云芳, 李延超. 2021. DS-YOLO: 一种部署在无人机终端上的小目标实时检测算法. *南京邮电大学学报(自然科学版)*, 41(1): 86-98) [DOI: 10.14132/j. cnki. 1673-5439.2021.01.011]
- Zhang X S, Wan F, Liu C, Ji R R and Ye Q X. 2019. FreeAnchor: learning to match anchors for visual object detection//Proceedings of the 22nd Conference on Neural Information Processing Systems. Vancouver, Canada; Curran Associates Inc.: 147-155
- Zhang Z S, Qiao S Y, Xie C H, Shen W, Wang B and Yuille A L. 2018c. Single-shot object detection with enriched semantics//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 5813-5821 [DOI: 10.1109/CVPR.2018.00609]

- Zhao G M, Ge W F and Yu Y Z. 2021. GraphFPN: graph feature pyramid network for object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 2763-2772 [DOI: 10.1109/ICCV48922.2021.00276]
- Zhao Q J, Sheng T, Wang Y T, Tang Z, Chen Y, Cai L and Ling H B. 2019. M2Det: a single-shot object detector based on multi-level feature pyramid network//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 9259-9266 [DOI: 10.1609/aaai.v33i01.33019259]
- Zhao Y Q, Rao Y, Dong S P and Zhang J Y. 2020. Survey on deep learning object detection. Journal of Image and Graphics, 25(4): 629-654 (赵永强, 饶元, 董世鹏, 张君毅. 2020. 深度学习目标检测方法综述. 中国图象图形学报, 25(4): 629-654) [DOI: 10.11834/jig.190307]
- Zheng M H, Gao P, Zhang R R, Li K C, Wang X G, Li H S and Dong H. 2020a. End-to-end object detection with adaptive clustering transformer [EB/OL]. [2022-01-18]. <https://arxiv.org/pdf/2011.09315.pdf>
- Zheng Z H, Wang P, Liu W, Li J Z, Ye R G and Ren D W. 2020b. Distance-IOU Loss: faster and better learning for bounding box regression//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12993-13000 [DOI: 10.1609/aaai.v34i07.6999]
- Zhong Q Y, Li C, Zhang Y Y, Xie D, Yang S C and Pu S L. 2020a. Cascade region proposal and global context for deep object detection. Neurocomputing, 395: 170-177 [DOI: 10.1016/j.neucom.2017.12.070]
- Zhong Z, Zheng L, Kang G L, Li S Z and Yang Y. 2020b. Random erasing data augmentation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 13001-13008
- Zhou D Z, Zhou X C, Zhang H W, Yi S and Ouyang W L. 2020. Cheaper pre-training lunch: an efficient paradigm for object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 258-274 [DOI: 10.1007/978-3-030-58598-3_16]
- Zhou P, Ni B B, Geng C, Hu J G and Xu Y. 2018. Scale-transferrable object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 528-537 [DOI: 10.1109/CVPR.2018.00062]
- Zhou X Y, Wang D Q and Krähenbühl P. 2019a. Objects as points [EB/OL]. [2022-01-18]. <https://arxiv.org/pdf/1904.07850.pdf>
- Zhou X Y, Zhuo J C and Krähenbühl P. 2019b. Bottom-up object detection by grouping extreme and center points//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 850-859 [DOI: 10.1109/CVPR.2019.00094]

- Zhu C C, He Y H and Savvides M. 2019. Feature selective anchor-free module for single-shot object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 840-849 [DOI: 10.1109/CVPR.2019.00093]
- Zhu X Z, Su W J, Lu L W, Li B, Wang X G and Dai J F. 2021. Deformable DETR: deformable transformers for end-to-end object detection//Proceedings of 2021 International Conference on Learning Representations. [s. l.]: OpenReview. net
- Zhu Y S, Zhao C Y, Wang J Q, Zhao X, Wu Y and Lu H Q. 2017. CoupleNet: coupling global structure with local parts for object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4146-4154 [DOI: 10.1109/ICCV.2017.444]

作者简介



曹家乐, 1991年生, 男, 副研究员, 主要研究方向为目标检测、语义分割、深度学习。

E-mail: connor@tju.edu.cn



庞彦伟, 通信作者, 男, 教授, 主要研究方向为目标检测、图像增强、医学影像分析、深度学习。

E-mail: pyw@tju.edu.cn

李亚利, 女, 助理研究员, 主要研究方向为目标检测、场景理解、视觉认知。E-mail: liyali13@tsinghua.edu.cn

孙汉卿, 男, 博士研究生, 主要研究方向为目标检测、双目视觉、深度学习。E-mail: HQSun@tju.edu.cn

谢今, 男, 副教授, 主要研究方向为目标检测、图像增强、深度学习。E-mail: jinjie@tju.edu.cn

黄凯奇, 男, 研究员, 主要研究方向为计算机视觉、模式识别、智能视觉监控、人的认知信息处理。

E-mail: kaiqi.huang@ia.ac.cn