

基于空洞卷积和特征融合的单阶段目标检测算法

李娟娟^{1,2}, 侯志强^{1,2}, 白玉^{1,2}, 程环环^{1,2}, 马素刚^{1,2}, 余旺盛³, 蒲磊⁴

(1. 西安邮电大学计算机学院, 西安, 710121; 2. 西安邮电大学陕西省网络数据分析与智能处理重点实验室, 西安, 710121;
3. 空军工程大学信息与导航学院, 西安, 710077; 4. 火箭军工程大学作战保障学院, 西安, 710025)

摘要 针对SSD多尺度目标检测过程中存在的目标漏检和错检问题,提出了一种融入多维空洞卷积和多尺度特征融合的目标检测算法。在卷积神经网络输出的多尺度特征中,浅层具有更多的细节信息,深层具有更多的语义信息,根据这一特点,对浅层网络采用了3种多维空洞卷积的浅层特征增强模块,获得具有语义信息的特征图,将增强后的特征图进行下采样,融合不同层的特征;同时在深层网络引入通道注意力模块,对通道进行权重分配,抑制无用信息,提高目标的检测性能。研究表明:该算法在PASCAL VOC数据集上检测精度为79.7%,比SSD算法提高了2.4%;在KITTI数据集上检测精度为68.5%,比SSD算法提高了5.1%,检测速度达到了实时性的要求,有效地改善了目标的漏检和错检。

关键词 目标检测;SSD算法;空洞卷积;特征融合

DOI 10.3969/j.issn.1009-3516.2022.01.014

中图分类号 TP391.4 **文献标志码** A **文章编号** 1009-3516(2022)01-0097-07

Single-Stage Object Detection Algorithm Based on Dilated Convolution and Feature Fusion

LI Juanjuan^{1,2}, HOU Zhiqiang^{1,2}, BAI Yu^{1,2}, CHENG Huanhuan^{1,2},
MA Sugang^{1,2}, YU Wangsheng³, PU Lei⁴

(1. College of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; 2. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; 3. Information and Navigation College, Air Force Engineering University, Xi'an 710077, China; 4. School of Operational Support, Rocket Force Engineering University, Xi'an 710025, China)

Abstract Aiming at the problem of missed detection and wrong detection in the process of SSD (Single Shot Multibox Detector) multi-scale object detection, this paper proposes a object detection algorithm that incorporates multi-dimensional dilated convolution and multi-scale feature fusion. Among the multi-scale features output by the convolutional neural network, the shallow layer has more detailed information, and the deep layer has more semantic information. According to this feature, this paper adopt three types of multi-dimensional dilated convolution shallow layers for the shallow network. The feature enhancement module obtains a feature map with semantic information, down-samples the enhanced feature map, and

收稿日期: 2021-05-19

基金项目: 国家自然科学基金(62072370)

作者简介: 李娟娟(1994—),女,陕西咸阳市人,硕士生,研究方向为目标检测。E-mail:15091421802@163.com

引用格式: 李娟娟,侯志强,白玉,等.基于空洞卷积和特征融合的单阶段目标检测算法[J].空军工程大学学报(自然科学版),2022,23(1):97-103. LI Juanjuan, HOU Zhiqiang, BAI Yu, et al. Single-Stage Object Detection Algorithm Based on Dilated Convolution and Feature Fusion [J]. Journal of Air Force Engineering University (Natural Science Edition), 2022, 23(1): 97-103.

fusion features of different layers; at the same time, a channel attention module is introduced in the deep network to assign weights to channels, suppress useless information, and improve object The detection performance. The research results show that the detection accuracy of the algorithm in this paper is 79.7% on the PASCAL VOC dataset, which is 2.4% higher than the SSD algorithm; the detection accuracy on the KITTI dataset is 68.5%, which is 5.1% higher than the SSD algorithm, and the detection speed reaches real-time requirements have effectively improved the missed and wrongly detected object.

Key words object detection; SSD algorithm; dilated convolution; feature fusion

目标检测作为计算机视觉领域的一个热门研究课题,在行人识别^[1]、自动驾驶^[2]和人脸检测^[3]等诸多方面具有广泛应用,近年来得到了迅速的发展。

传统的目标检测方法通常利用手工特征进行检测,如 HOG^[4] (histograms of oriented gradient) 和 SIFT^[5] (scale-invariant features transfer) 等特征,但是这些特征处理过程复杂,检测速度慢,在对不同场景的目标进行检测时,算法的鲁棒性不高。基于深度学习的目标检测算法根据有无建议区域生成阶段分为单阶段(one stage)目标检测算法和两阶段(two stage)目标检测算法;单阶段检测算法主要有 SSD^[6] 和 YOLO 系列^[7-10] 等;两阶段检测算法主要有 R-CNN^[11]、SPP-Net^[12] 和 Faster R-CNN^[13] 等。由于有锚框的检测算法在检测时会带来正负样本不均衡等问题,近年来出现了无锚框(anchor-free)的目标检测算法,主要算法有 CornerNet^[14] 和 CenterNet^[15] 等。CornerNet 利用了左上角和右下角的两个关键点来进行目标检测;CenterNet 将目标作为点的模型,只检测目标的中心点,通过预测目标的中心点直接回归对应目标的长和宽。

在上述算法中,单阶段 SSD 算法的精度和两阶段 Faster R-CNN 算法的精度比较接近,但速度达

到了实时性的要求,是目标检测领域一个经典的算法。因此,许多研究人员对 SSD 算法进行了改进,改进的算法有 DSSD 算法^[16]、RSSD 算法^[17]、RefineDet 算法^[18]、RFBNet 算法^[19] 和 FSSD 算法^[20] 等。

SSD 算法克服了 Faster R-CNN 和 YOLOv1 算法的不足,提出了多尺度目标检测,考虑了大目标和小目标,平衡了精度和速度。但是也存在一定的问题:SSD 在对多尺度目标进行检测时,使用浅层网络检测小目标,深层网络检测大目标,小目标往往更依赖浅层特征的空间信息,因为浅层特征有更高的分辨率,但是小目标在浅层表征力不足,缺乏语义信息,导致分类比较困难。

针对 SSD 在目标检测过程中存在的目标漏检和误检的问题,本文提出了一种融入多维空洞卷积和多尺度特征融合的目标检测算法。

1 基于空洞卷积和特征融合的单阶段目标检测算法

1.1 算法框架

本文算法框架见图 1。

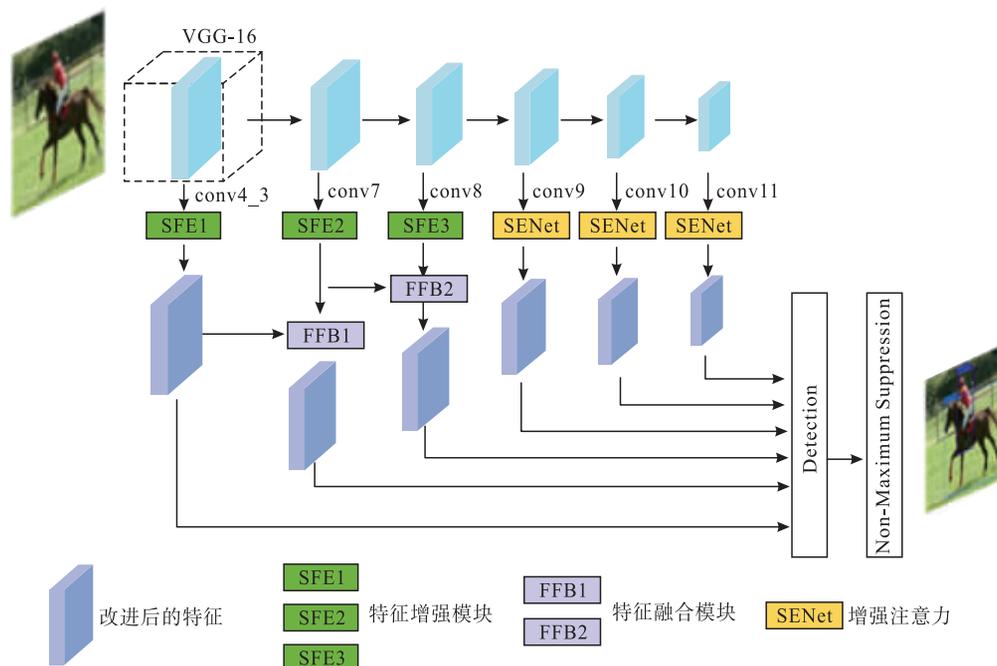


图 1 本文算法框架

本文算法是以 SSD 目标检测算法为基础,将特征层分为浅层和深层两部分,浅层分别为:conv4_3 层、conv7 层和 conv8 层,深层分别为:conv9 层、conv10 层和 conv11 层。浅层分辨率高,具有丰富的细节信息,但特征表征力不足^[16];深层分辨率低,具有丰富语义信息,能很好地表示中等目标和大目标,基于此,对浅层特征和深层特征分别进行改进。在浅层引入 3 个浅层特征增强模块 SFE 和 2 个特征融合模块(feature fusion block,FFB),处理 conv4_3、conv7 和 conv8 层的特征,3 个浅层特征增强模块使用不同维的空洞卷积,以扩大特征的感受野;FFB1 模块融合特征增强后的 conv4_3 和 conv7,FFB2 模块融合特征增强后的 conv7 和 conv8。对于 conv9 层、conv10 层和 conv11 层引入通道注意力模块,使网络在全局范围内选择更有用的目标信息,增强了特征的表达能力。下面对基准算法、特征增强模块、特征融合模块和通道注意力展开说明。

1.2 基准算法

SSD 算法的特征提取网络是 VGG16,SSD 算法使用 VGG16 的 conv4_3 层,和将全连接 fc7 层转化成的 conv7 层做检测,并在 VGG16 基础上,进一步

延伸了 4 个卷积层 conv8、conv9、conv10 和 conv11 层用来做检测,所以 SSD 算法共使用 6 个不同尺度的特征图进行分类与回归预测。

1.3 浅层特征增强模块

在深度网络中,通常通过下采样操作来增加特征图感受野,但感受野增加的同时也造成特征图分辨率降低,细节信息会减少。空洞卷积可以增加特征图的感受野,同时也不会损失图像信息。许多研究人员在网络中使用不同方式的空洞卷积^[21],使性能有一定的提升。图 2 分别给出了普通卷积和空洞卷积膨胀率为 2 的示意图。

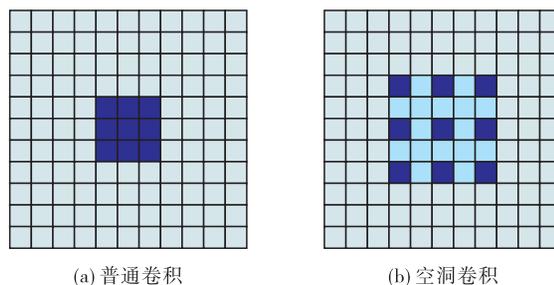


图 2 普通卷积和空洞卷积示意图
本文使用 SFE 对特征图扩大感受野,见图 3。

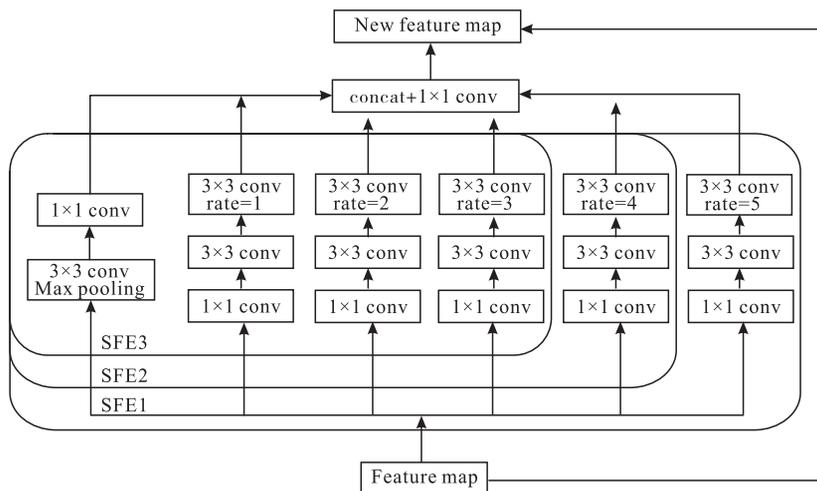


图 3 浅层特征增强模块

设计了 SFE1、SFE2 和 SFE3 这 3 个具有多维空洞卷积的浅层特征增强模块,对 conv4_3 层、conv7 层和 conv8 层进行感受野扩大。本文浅层特征增强模块采用残差网络结构,针对现有的 Inception 结构进行改进,将不同膨胀率的空洞卷积代替原有卷积层,形成多分支空洞卷积。本文使用 5 种不同膨胀率(rate=1,2,3,4,5)的空洞卷积来对原始的特征提取信息,获得 5 种不同语义信息的特征图,使用不同膨胀率得到的特征图进行融合,可以提取图像中不同大小的目标信息,膨胀率较小的空洞

卷积可以提取图像中的小目标信息,有利于小目标的准确分类,增强了模型对于小目标的检测能力。

目标的特征信息会随着网络的传播由浅层的细节信息转化为深层的语义信息,特征的感受野也会增大,且特征每层的尺寸是不同的,所以膨胀率的选择应该与特征层尺寸大小和所具有的信息相关,才能保证目标可以被检测到,不会造成丢失。通过实验对比选择不同浅层特征增强模块和相同的浅层特征增强模块的检测结果,根据特征尺寸和感受野来选择不同维的空洞卷积,对于中小目标的检测较好

一些。

下面对 SFE1、SFE2 和 SFE3 模块分别进行说明:在 conv4_3 层添加 SFE1 模块,使用 1×1 卷积将原始特征的通道数由 C 变为 $C/6$ 赋给每个分支,使用膨胀率为 1,2,3,4,5 的空洞卷积对原始特征进行处理。在 conv7 层添加 SFE2 模块,使用 1×1 卷积将原始特征的通道数由 C 变为 $C/5$ 赋给每个分支,使用膨胀率为 1,2,3,4 的空洞卷积。在 conv8 层添加 SFE3 模块,使用 1×1 卷积将原始特征的通道数由 C 变为 $C/4$ 赋给每个分支,使用膨胀率为 1,2,3 的空洞卷积。将各个分支按通道拼接融合,然后将原始特征和经过多个不同膨胀率的空洞卷积的特征进行相加融合,保留原始特征信息。

1.4 特征融合模块

不同的特征层经过特征增强后的信息是不同的,将不同分辨率并行的特征层进行融合,提高了网络的表征能力,增强了小目标的特征表达。常用的特征融合方式有相加(add)和拼接(concat),本文选用的特征融合方式是拼接,这样融合特征通道具有一定的灵活性。FFB1 模块融合的是 SFE1 输出特征图经过 stride 为 2 的下采样变成 19×19 的特征图,和当前 SFE2 输出的特征图;FFB2 模块融合的是 SFE2 输出的特征图经过 stride 为 2 的下采样变成 10×10 的特征图,和当前 SFE3 输出的特征图。特征融合模块(FFB)结构如图 4 所示。

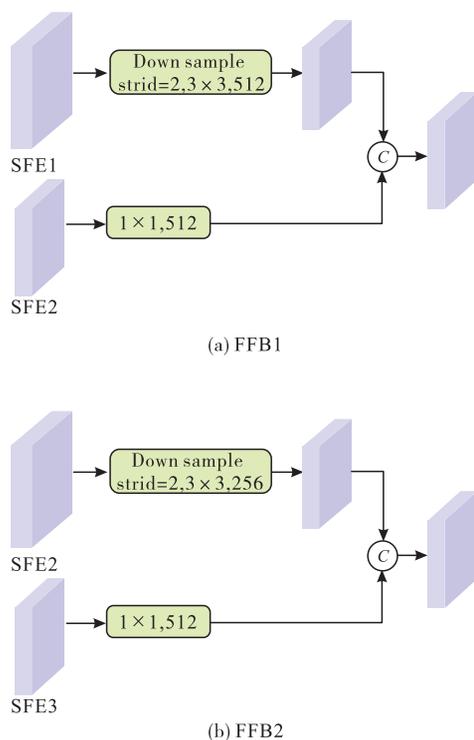


图 4 特征融合模块

1.5 通道注意力模块

通道注意力是通过特征通道进行权重再配来提高特征表达能力。因为原始特征图每个通道占比都是一样的,所以在输出检测时,不能聚焦于重点信息,而是全局搜索,这样会消耗训练时间,且很难找到重点的部分。对特征使用通道注意力,能在图像中更容易找到重点关注的区域,从而更好检测到目标。本文采用的通道注意力模块是 SENet^[22],由压缩(squeeze)、激励(excitaion)和权重重标定(re-weight)3个部分组成。

2 实验结果与分析

2.1 实验设备和数据集

本文算法使用的实验平台为 Ubuntu16.04 系统,GPU 为 1080Ti,深度学习框架为 PyTorch1.2。实验使用 PASCAL VOC 数据集和 KITTI 数据集。

2.2 实验评价指标

为了更准确地衡量目标检测的质量,本文使用的指标为平均精度均值(mean average precision, mAP)和每秒帧率(FPS)来衡量目标检测质量,mAP 是指所有类别的精度均值。

2.3 定性分析

为了评估本文算法的检测性能,在 PASCAL VOC 数据集和 KITTI 数据集上将 SSD 算法的检测效果图与本文算法的检测效果图进行对比。

2.3.1 PASCAL VOC 数据

在 PASCAL VOC 数据集上检测结果如图 5 所示,从第 1 列可以看出,图像中被遮挡的目标“cow”,SSD 算法没有检测到,但是本文算法能检测到,并且分类分数提高了。从第 2 列可以看出“person”SSD 算法没有检测到,但是本文算法可以检测到。从第 3 列可以看出 SSD 算法将石头误检成“sheep”也有漏检情况,本文算法对目标的漏检和误检有一定的改善。从第 4 列可以看出,图像中的“pottedplant”没有被检测到,本文算法能检测到,“sofa”的分类分数提高了。从图 5 检测效果图可以看出,在图像中存在多目标的情况下,本文算法也可以检测到,并且不会出现误检和漏检的情况,且分类分数都有一定的提升,有效地改善了多目标的检测效果。

2.3.2 KITTI 数据集

在 KITTI 数据集上检测结果如图 6 所示,在第 1 列中 SSD 算法没有检测到目标“pedestrian”,而本文算法能检测到。在第 2 列 SSD 算法将目标“car”漏检了,而本文算法则能检测到,“car”的分类分数也提高了。本文算法对漏检有一定的改善。



图 5 在 PASCAL VOC 数据集上实验结果对比

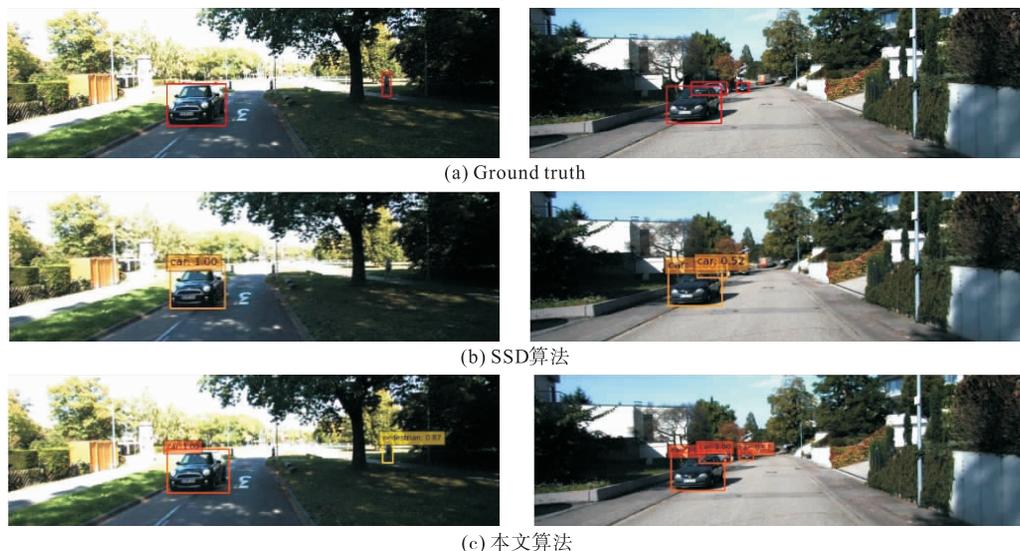


图 6 在 KITTI 数据集上实验结果对比

2.4 定量分析

2.4.1 PASCAL VOC 数据集

在 PASCAL VOC 数据集上, 本文算法与两阶段 Faster R-CNN 算法^[13]和 R-FCN 算法^[23]、单阶段 YOLOv3 算法^[9]、CenterNet 算法^[15]以及 SSD 改进算法, DSSD 算法^[16]、RSSD 算法^[17]和 FSSD 算法^[20]等算法检测结果对比, 如表 1 所示。

从表 1 中可以看出, 本文算法的精度为 79.7%, 和原始 SSD 算法及其改进算法进行比较, 本文算法的检测精度最高。比原始 SSD 算法提高了 2.4%, 相比 DSSD 算法、RSSD 算法、FSSD 算法, 本文算法分别提高了 1.1%、1.2%、0.9%, 相比文献[24]提高了 1.6%, 在速度方面原始的 SSD 算法为 42FPS, 改进算法的速度为 34FPS, 达到了实时性的要求。和单阶段经典算法 YOLOv3 相比提高了 0.4%, 和无锚框算法 CenterNet 相比提高了

1.0%。和两阶段算法 Faster R-CNN、R-FCN 相比, 分别提高了 6.5% 和 0.2%, 速度相比两阶段算法达到了实时性的要求。

表 1 PASCAL VOC 数据集上检测结果对比

Method	Backbone	mAP	FPS
SSD	VGG-16	77.3	42
Faster R-CNN	VGG-16	73.2	7
R-FCN	ResNet-101	79.5	5.8
YOLOv3	DarkNet-53	79.3	26
CenterNet	ResNet-101	78.7	30
DSSD	ResNet-101	78.6	14
RSSD	VGG-16	78.5	35
FSSD	VGG-16	78.8	66
文献[24]	VGG-16	78.1	39
Ours	VGG-16	79.7	34

2.4.2 KITTI 数据集

在 KITTI 数据集上本文算法与 SSD 算法检测

结果进行对比,如表2所示。

表2中可以看出,本文算法的检测精度为68.5%,比SSD算法提升了5.1%,单个类别“car”提升了3.2%，“cyclist”提升了8.4%，“pedestrian”提升了3.7%,FPS为45,满足实时性检测的要求。

表2 KITTI数据集上检测结果对比

Method	car	cyclist	pedestrian	mAP	FPS
SSD	85.3	57.6	47.3	63.4	56
Ours	88.5	66.0	51.0	68.5	45

2.5 消融实验

通过添加单个模块和组合多个模块测试,说明每个模块在改进算法中的作用,在PASCAL VOC上实验结果如表3所示。

表3 消融实验

SSD	SFE	FFB	SENet	mAP
✓				77.3
✓	✓			78.9
✓	✓	✓		79.5
✓	✓	✓	✓	79.7

从表3中可以看出,将SFE模块添加到SSD算法中,提高了原始算法的精度,mAP由77.3%提升到78.9%,提高了1.6%,证明了扩大感受野对于浅层特征图的重要性。在加入SFE模块的基础上加入特征融合FFB模块,提高了0.6%,证明了浅层特征信息对较深一层的的信息是有补充作用的。在此基础上添加了通道注意力模块,提高了0.2%,对特征通道进行筛选,强调了重要的通道信息。

3 结语

在SSD目标检测算法的基础上,针对不同层信息的差异,本文首先在特征图上添加3个不同的多维空洞卷积SFE模块;其次使用2个特征融合模块融合不同层的特征,使特征层之间的信息可以互补,更好的检测到目标;最后在网络深层添加通道注意力模块,对特征进行加权处理,强调有用的信息,增强特征的可分辨性。实验结果表明,所提出算法有效地改善了目标漏检和错检,在检测速度满足实时性的情况下,提升了目标的检测精度。

参考文献:

[1] TANAKA S, INOUE Y. Outdoor Human Detection with Stereo Omnidirectional Cameras[J]. Journal of Robotics and Mechatronics, 2020, 32(6): 1193-1199.
 [2] LI P, CHEN X, SHEN S. Stereo R-CNN Based 3D Object Detection for Rautonomous Driving[C]//Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]:IEEE, 2019: 7644-7652.
 [3] CHAVES D, FIDALGO E, ALEGRE E, et al. Assessment and Estimation of Face Detection Performance Based on Deep Learning for Forensic Applications[J]. Sensors, 2020, 20(16): 4491.
 [4] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005, 1: 886-893.
 [5] LOWE D G. Object Recognition from Local Scale Invariant Features [C]//Proceedings of the Seventh IEEE International Conference on Computer Vision. [S. l.]:IEEE, 1999: 1150-1157.
 [6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot Multibox Detector[C]//European Conference on Computer Vision. [S. l.]: Springer, Cham, 2016: 21-37.
 [7] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2016: 779-788.
 [8] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]:IEEE, 2017: 7263-7271.
 [9] REDMON J, FARHADI A. Yolov3: An Incremental Improvement[J]. ArXiv Preprint ArXiv:1804.02767, 2018.
 [10] BOCHKOVSKIY A, WANG C Y, LIAO H Y M, et al. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. ArXiv Preprint ArXiv:2004.10934, 2020.
 [11] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]:IEEE, 2014: 580-587.
 [12] HE K, ZHANG X, REN S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.
 [13] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region on Neural Information Processing Systems[R] Massachusetts: MIT Press, 2015:91-99.
 [14] LAW H, DENG J. Cornernet: Detecting Objects as Paired Keypoints[C]//Proceedings of the European

