

基于注意力机制的遥感图像车辆目标检测

肖美波,孟维伟,刘轩宏,张丹[†]

(南京林业大学,江苏 南京 210037)

摘要:为了更好地通过遥感图像监测地面交通流量,应对日趋复杂的路况问题,提出了一种基于YOLOX的全新的一阶段目标检测网络 AE-YOLO(Attention-Enhanced YOLO)。采用混合注意力加强的SCACSP特征增强模块,很大程度上抑制了背景信息,突出了目标的显著特征,用解耦头部代替耦合头,大大提高了网络的收敛速度和性能。对损失函数进行改进,使用CIoU Loss代替原来的IoU Loss,从而解决了树木遮挡等造成的车辆漏检问题。试验结果表明,与主流的目标检测网络相比,该方法在精度上有更好的表现。

关键词:深度学习;目标检测;遥感图像;混合注意力机制;YOLOX

中图分类号:TP39

文献标识码:A

Vehicle Target Detection in Remote Sensing Image Based on Attention Mechanism

XIAO Meibo, MENG Weiwei, LIU Xuanhong, ZHANG Dan[†]

(Nanjing Forestry University, Nanjing, Jiangsu 210037, China)

Abstract: In order to better monitor ground traffic flow through remote sensing images and deal with increasingly complex road conditions, a new one-stage target detection network AE-YOLO (Attention-Enhanced YOLO) based on YOLOX is proposed. The SCACSP feature enhancement module with mixed attention enhancement suppresses the background information to a large extent, highlights the salient features of the target. Replacing the coupling head with the decoupled head, which greatly improves the convergence speed and performance of the network. Using CIoU Loss instead of the original IoU Loss to improve the loss function, thus solving the problem of vehicle missed detection caused by trees occlusion and other reasons. Experimental results show that the proposed method outperforms the mainstream object detection networks in accuracy.

Key words: deep learning; target detection; remote sensing image; hybrid attention mechanism; YOLOX

近年来,遥感影像的地面分辨率不断提高,研究者们开始对地面微小目标的探测进行研究。遥感图像的优点是空间上城市级别范围的连续覆盖,这是摄像头和移动街景等其他交通监测方式所不具备的。所以从高分辨率遥感图像上进行车辆信息提取,可以作为地面交通流量监测的一种非常重

要的补充方式。城市公路交通系统压力日益增大,需要更加智能的方式来应对日趋复杂的路况问题。因此,如何利用基于深度学习的计算机视觉技术来监测车流量变得至关重要。

本文提出了一个基于YOLOX^[1]的全新的一阶段目标检测网络 AE-YOLO (Attention-En-

hanced YOLO),以解决遥感图像中车辆目标物小、信息有限,以及由树木遮挡引起的漏检问题。

1 混合注意力模块增强的目标检测网络
AE-YOLO

目前遥感图像车辆检测的方法主要分为两类,分别是传统方法和基于深度学习的方法。传统方法大多需要较多的人为干预,成本较高,并且在一些较复杂场景中检测精度也无法满足使用要求^[2]。基于深度学习的方法可以分为两类,分别是两阶段算法和一阶段算法^[3,4]。两阶段算法中最早的是 R-CNN^[5]算法,而 Fast R-CNN^[6]和 Faster R-CNN^[7]算法都是对此进行改进,其特点是精度高,但是速度相对较慢。一阶段算法中最有代表性的是 YOLO^[8]系列算法,相对于两阶段算法,其算法速度快,但精度稍逊,是一个速度和精度相对均衡的算法。基于 YOLO 算法,YOLO9000^[9],YOLOv3^[10],YOLOv4^[11],YOLOv5^[12]等一系列算法也被研发出来。

由于遥感图像的特殊性,并且 anchor-based 的方法需要设定额外的超参,因此会产生大量多余的冗余框(完全不包含目标的背景框),额外增加不必要的计算量。有鉴于 FA-YOLO^[13]模型以及 YOLOv5s-THSE^[14]模型都是基于 YOLO 系列算法,分别在 YOLOv4 和 YOLOv5s 模型中加入不同的注意力模块,提高了复杂背景下的小目标检测能力。为了克服这些问题,同时由于 YOLOX 模型网络更容易收敛且准确度更高,因此提出了基于 YOLOX 为 baseline 的单阶段 anchor-free 实时性目标检测框架 AE-YOLO,针对小目标检测提出了混合注意力加强的 SCACSP 特征增强模块,提高遥感图像中小目标车辆的显著特征信息,以提高分类和定位的精度。并且融入多尺度特征模块,解决了在遥感图像中车辆小目标难以检测到的问题^[15,16]。最后,为了加快收敛速度,使用解耦头代替原来的耦合头部。目标检测网络 AE-YOLO 的总体结构如图 1 所示,网络主要包括三个部分:backbone 主干特征提取网络、aeapan 特征加强模块和改进的解耦检测头。

1.1 主干特征提取网络

主干特征提取网络是在 CSPDarknet53 的基础上改进的。采用了残差结构(residual structure),CSPDarknet 中的残差卷积可以分为两个部

分,主干部分是一次的卷积和一次的卷积;残差边不作任何处理,直接将主干的输入与输出结合。整个 AE-YOLO 的主干部分都由残差卷积构成。残差网络的优点在于容易优化,并且能够通过增加相当的深度提高网络预测的准确率。其内部的残差块使用了跳跃连接,缓解了在深度神经网络中增加深度带来的梯度消失问题。

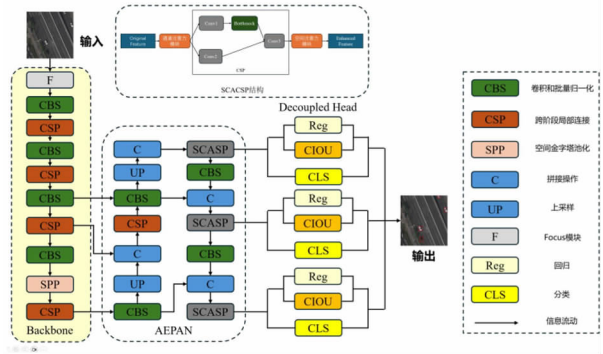


图 1 网络总体结构图

在 AE-YOLO 中的主干特征提取网络中使用跨阶段局部增强网络 CSPNet^[17],从而提高特征提取网络的特征抽取能力,同时也可以显著降低计算瓶颈和大量减少内存流量。CSPNet 结构是将原来堆叠的残差块进行了一个拆分,拆成了左右两部分:主干部分继续进行残差块的堆叠,另一部分则在经过了一次卷积之后,直接连接到最后。因此可以认为在 CSP 中存在一条较大的残差边。

在主干提取网络的最开始采用了 Focus 模块。Focus 模块对图片进行切片操作,在图像中每隔一个像素取一个值,得到了四个独立的特征层,然后将四个独立的特征层进行堆叠,此时宽高信息就集中到了通道信息,输入通道扩充了四倍,拼接起来的特征层相对于原先的三通道变成了十二通道。这种方式可以避免由下采样导致的信息损失。

激活函数使用的是 SiLU,其为 Sigmoid 和 ReLU 的改进版,在深层模型上的效果要优于 ReLU。

在 AE-YOLO 的主干特征提取网络中也使用了特征金字塔模块(SPP^[18]),通过不同池化核大小的最大池化进行特征提取,增加网络的感受野。同时将 SPP 模块嵌入主干特征提取网络中,加强了主干特征网络对多尺度信息的敏感程度。

1.2 混合注意力加强引导的 PAN 特征加强提取网络

采用自上而下和自下而上的多尺度聚合网络 PAN 来融合浅层和深层特征图特征去实现平衡分

类和定位。由于遥感图像中背景信息和主体车辆目标比较相似,目标与背景的区分度不高,而传统的 PAN 没有考虑从目标中提取到的重要信息,常常导致背景被误检测为目标车辆,为了解决上述问题,本文提出了混合注意力加强引导的多尺度聚合特征加强提取网络 EAPAN,其核心模块 SCACSP 结构如图 2 所示。

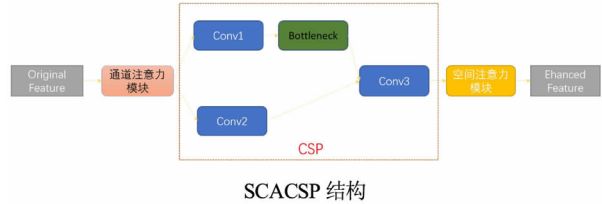


图 2 SCACSP 结构图

全新的混合注意力模块很大程度上抑制了背景信息,突出了目标的显著特征。其大致可以分为三个子模块:通道注意力模块、跨阶段局部特征提取(CSP)模块和空间注意力模块。SCACSP 结构的工作原理可以表示为:

$$C_s = CA(c) \quad (1)$$

$$C_p = CSP(C_s) \quad (2)$$

$$P_c = SA(C_p) \quad (3)$$

其中, s 是从主干特征网络中提取出的特征,SA 和 CA 分别是空间注意力模块和通道注意力模块,CSP 是跨阶段局部特征提取模块。

主干特征网络提取到的特征 c 首先通过通道注意力模块,通道注意力模块对输入特征图的不同通道进行区分和不同的加权,以突出通道的不同重要性。为了防止过拟合,并且使得模型尽量轻量化,提高实时性检测的能力,选用了 SE 模块作为通道注意力进行加强特征提取。通道注意力模块结构如图 3 所示。

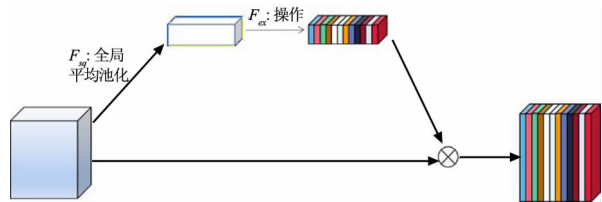


图 3 通道注意力模块结构图

SE 模块^[19]的操作主要分为三步:压缩操作、激励操作和 scale 操作。

步骤一:输入特征图通过压缩操作,即全局平均池化操作,将特征图映射到 $1 \times 1 \times c$ 的向量。

步骤二:使用激励操作,用两个大小为 1×1 的

卷积核来学习每个通道的权重系数。

步骤三:使用 scale 操作将原有的特征乘以通道权重(0,1),从而得出输出结果 $W \times H \times C$ 的特征图。

通道注意力的操作流程可以表示为:

$$Z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (4)$$

$$S_c = F_{ex}(W, z_c) = \sigma(W_2 \delta(W_1, z_c)) \quad (5)$$

$$\tilde{X}_c = s_c \times u_c \quad (6)$$

其中, F_{sq} 是压缩操作, F_{ex} 是激励操作。

特征的空间信息直接决定了目标的定位准确性。为了抑制背景信息并显著突出目标的特征信息,在 CSP 模块之后增加了空间注意力模块,如图 4 所示。

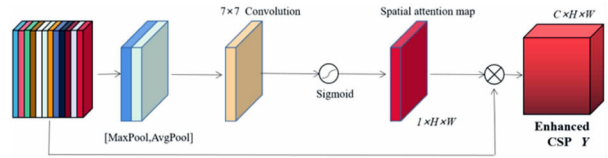


图 4 空间注意力模块

将 CSP 模块输出的特征图作为本模块的输入特征图。首先对特征图做基于通道的全局平均池化和全局最大池化操作得到两个 $W \times H \times 1$ 的特征图,然后将这两张特征图作拼接操作。经过一个卷积操作,降维为 1 个通道。再通过一个 Sigmoid 激活函数生成空间注意力特征图,最后将该特征图和原来的特征输入作点乘,得到最终生成的特征。空间注意力的操作流程可以表示为:

$$M_s(F) = \sigma(\text{Conv}(F_{\text{avg}}^s; F_{\text{max}}^s)) \quad (7)$$

其中, F_{avg}^s 表示全局平均池化, F_{max}^s 表示全局最大池化,Conv 是卷积操作, σ 表示 Sigmoid 激活函数。

1.3 解耦检测头

在本文提出的 AE-YOLO 中,使用解耦头(decoupled head)代替原来的耦合头部。解耦头考虑到分类和定位所关注的内容并不相同,因此将分类和定位分离,采用不同的操作,分类用全连接层,定位用卷积层。解耦头的使用,大大提高了网络的收敛速度和性能。

2 改进的损失函数 CIOU Loss

采用 CIOU Loss 代替原来的 IoU Loss,解决了树木遮挡等造成的车辆漏检问题。

IoU Loss 具有尺度不变性,不论 box 的尺度

大小,输出的损失值一定在 0~1 之间,从而对形状和尺度具有很强的鲁棒性。然而,当 IoU Loss 用在遥感图像检测中时,由于树木的遮挡和背景与目标物体的颜色比较一致,基于 IoU 的损失函数往往表现不佳,导致检测精度下降。

为了解决上述问题,引入了 CIoU Loss^[20] 来计算边界框回归损失。CIoU Loss 认为两个矩形框相对比例的一致性也是一个重要的几何因素,因此 CIoU Loss 的惩罚项把预测框的纵横比拟合真实框的纵横比考虑进去了。此外,当预测框与真实框重叠时,直接最小化两个框之间的距离,使其收敛速度快于 IoU Loss。

CIoU Loss 的惩罚项定义如下:

$$R_{ciou} = \frac{\rho^2(b, b^{gt})}{c^2} + \alpha$$

(8)

其中 α 用来做平衡的参数, v 是用来衡量长宽比一致性的参数, ρ 是两个矩形框之间的欧氏距离, b 和 b^{gt} 表示两个矩形框的中心点, c 表示两个矩形框闭包区域的对角线的距离。完整的 CIoU Loss 的公式如下:

$$L_{ciou} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha$$

(9)

3 实验结果

3.1 数据集

为了证明模型在不同数据集上的通用性和鲁棒性,本文收集了两个高分辨率遥感图像数据集。第一个数据集包含 500 张左右图片,是与江苏测绘局合作获得的,取景的是玄武区道路交通状况。其背景和前景内容丰富,并且部分车辆被树木遮挡住。我们使用 Labelimg 进行标注,标注信息存放在 xml 文件中。另一个数据集由 Shen 等人^[21] 采集,也是在南京市上空捕获的。其中包含了 615 张高分辨率航拍图,分辨率为 1368×770 。无人机在 60 m 的高空俯拍到了这些图像,车辆信息被标注在每张图像上,每张图片大约有 30 种汽车的样本被注释。因此整个数据集大概包含大约 18450 个具有 ground truth 的样本。

3.2 实验环境/训练细节

本文的实验是在 Ubuntu 操作系统上进行的;其内存为 32 GB;GPU 为 Tesla V100;学习框架为 Pytorch。将两个数据集中的图片以 7:3 的比例随机分成训练集和测试集,并同时使用随机裁剪、随机翻转和对比度增强等手段进行数据增强。把

输入的图片统一缩放为 640 像素 $\times$ 640 像素。在训练时加载了在 VOC 数据集上训练好的权重。为了保护权重不被破坏并且加快训练,本文通过冻结特征提取层来训练前 50 个迭代(epoch),初始化学率为 0.0001,训练批次为 4。在后 50 个迭代中,解冻特征提取层以训练整个网络。在训练过程中使用了 Adam 优化器,将动量和权重衰减分别设置为 0.9 和 0.0005。

相关训练参数见表 1:

表 1 模型训练参数

冻结阶段(Freeze_Epoch)	50
解冻阶段(UnFreeze_Epoch)	50
初始学习率(Init_lr)	0.0001
训练批次(Batch)	4
优化器(Optimizer)	adam
动量(Momentum)	0.9
权重衰减(Weight_Decay)	0.0005

3.3 评价指标

本文采用目标检测模型常用评价指标 mAP 进行模型评估。AP 指 PR(Precision-Recall)曲线下的面积,mAP 指每个类别 AP 的均值。AP 以及 mAP 值越大越好。Precision 和 Recall 的计算公式如下:

$$Precosion = \frac{TP}{FP + TP}$$

(10)

$$Recall = \frac{TP}{TP + FN}$$

(11)

TP 表示被分为正样本,且分对的;FP 表示被分为正样本,但分错的;FN 表示被分为负样本,但分错的。

3.4 结果分析

本文所提出的 AE-YOLO 可视化结果如图 5 所示。

3.4.1 对比试验

为了验证本文提出模型 AE-YOLO 的有效性,将其与主流的目标检测模型 SSD^[22],Faster-RCNN,YOLOv3,YOLOv4 和 YOLOX-s 在两个数据集上进行对比实验,对比实验结果如表 2 所示。由表 2 分析可知,在两个数据集上 AE-YOLO 的 mAP 分别达到了 59.06% 和 96.76%,分别较原来的 YOLOX 提高了 1.06% 和 0.55%,相比其他主流的目标检测模型也有更好的检测性能。



图 5 AE-YOLO 检测效果图

表 2 主流目标检测模型对比结果

数据集	模型	Precision	Recall	F1	mAP
我们自己收集的 数据集	Faster-RCNN	50.06%	32.41%	0.393	39.82%
	SSD	63.51%	33.73%	0.441	45.81%
	YOLOv3	67.85%	38.42%	0.491	49.17%
	YOLOv4	65.33%	43.09%	0.519	52.92%
	YOLOX	69.58%	45.42%	0.550	58.01%
	AE-YOLO	70.57%	46.92%	0.564	59.06%
Nanjing 数据集	Faster-RCNN	77.65%	70.54%	0.739	75.12%
	SSD	85.16%	80.87%	0.828	82.43%
	YOLOv3	80.17%	86.59%	0.833	85.11%
	YOLOv4	87.21%	94.68%	0.908	93.51%
	YOLOX	87.43%	96.81%	0.919	96.21%
	AE-YOLO	87.99%	97.62%	0.926	96.76%

此外,我们同时也在较新的 YOLOv7^[23]网络上训练了 Nanjing 数据集,相关结果见表 3。

表 3 YOLOv7 在 Nanjing 数据集训练结果

Precision	Recall	F1	mAP
81.54%	88.58%	0.8491	89.10%

3.4.2 消融实验

本文提出的 AE-YOLO 中有两个关键模块 SCACSP 混合注意力模块和 CIoU Loss 损失函数。为了验证这两个模块的有效性,本文设计了消融实验。在 CSP 跨阶段局部特征提取模块中,本文引入了混合注意力模块,并使用 CIoU Loss 代替了原来的 IoU Loss。本文进行了四组消融实

验:(1)baseline 模型 YOLOX;(2)用 CSACSP 模块代替原来的 CSP 特征提取模块;(3)用 CIoU Loss 代替原来的 IoU Loss;(4)本文提出的注意力加强引导模型 AE-YOLO。

对比实验结果如表 4 所示。表 4 表明本文所提出的 AE-YOLO 在精度、召回率和 mAP 方面都有显著的提高。与起初的 YOLOX 进行比较,混合注意力模块的引入在两个数据集上分别增加了 0.88%和 0.2%的 mAP。而 CIoU 损失函数的使用在两个数据集上的提高分别为 0.72%和 0.37%。并且在训练期间,两个数据集上的 CIoU 损失值始终低于 IoU 损失。总之,AE-YOLO 相比于原始的 YOLOX,拥有更好的检测性能。

表 4 不同改进方法的实验结果

数据集	模型	Precision	Recall	F1	mAP
我们自己收集的 数据集	YOLOX	69.58%	45.42%	0.549	58.01%
	SCACSP	70.28%	45.51%	0.552	58.89%
	CIoU	69.72%	45.61%	0.551	58.73%
	AE-YOLO	70.57%	46.92%	0.563	59.06%
Nanjing 数据集	YOLOX	87.43%	96.81%	0.918	96.21%
	SCACSP	87.57%	96.93%	0.920	96.41%
	CIoU	88.13%	97.37%	0.925	96.58%
	AE-YOLO	87.99%	97.62%	0.925	96.76%

4 结 论

考虑到遥感图像中背景和前景相似度高,并且其中的车辆目标物极小,引入了混合注意力模块,提出了全新的 CSACSP 结构,从而能够提取出更具有辨别性的特征信息,提高分类和定位的精度。与此同时,为了解决由树木和车辆之间的遮挡引起的漏检误检问题,本文引入 CIoU Loss 代替原来的 IoU Loss,使得模型具有更强的鲁棒性,并且在训练过程中,CIoU Loss 相比于原来的损失函数收敛速度更快。实验结果表明,与主流的目标检测模型相比,本文提出的 AE-YOLO 具有更好的检测性能,有更好的鲁棒性。

受研究时间和数据规模的限制,目前的遥感图像车辆检测都是在白天光照条件比较好的情况下实现的,后续将通过域适应或者风格迁移等生成对抗网络的技术增强夜间的车辆特征,从而进一步验证模型在黑夜和光照度低的条件下对车辆目标的识别和检测。

参考文献

[1] GE Z, LIU S, WANG F, et al. YoloX: exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[2] 彭新月,张昊明,钟若飞. 改进 YOLOv3 模型的 GF-2 卫星影像车辆检测[J]. 测绘科学,2021,46(12):147-154.

[3] 邹伙宗,邓守城. 基于深度学习的车辆目标检测算法综述[J]. 时代汽车,2023(15):16-18.

[4] 包晓敏,王思琪. 基于深度学习的目标检测算法综述 [J]. 传感器与微系统, 2022,41(4):5-9.

[5] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014:580-587.

[6] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015:1440-1448.

[7] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems, 2015:28.

[8] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 779-788.

[9] REDMON J, FARHADI A. YOLO9000: better, faster,

stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017: 7263-7271.

[10] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[11] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[12] FANG Y, GUO X, CHEN K, et al. Accurate and automated detection of surface knots on sawn timbers using YOLO-V5 model[J]. BioResources, 2021, 16(3): 5390.

[13] DU S, ZHANG B, ZHANG P, et al. FA-YOLO: an improved YOLO model for infrared occlusion object detection under confusing background[J]. Wireless Communications and Mobile Computing, 2021, 2021(1):1-10.

[14] LIANG C, YAN Z, REN M, et al. Improved YOLOv5 infrared tank target detection method under ground background [J]. Scientific Reports, 2023, 13(1): 6269.

[15] 邱地发,于淑芳,刘锦辉等. 基于 YOLOv5s 的遥感图像的车辆小目标检测[J]. 计算机与现代化,2023(5):122-126.

[16] 张奇. 基于 YOLO 网络的小目标检测算法研究[D]. 绵阳:西南科技大学,2023.

[17] WANG C Y, LIAO H Y M, WU Y H et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops,2020: 390-391.

[18] 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.

[19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:7132-7141.

[20] ZHENG Z, WANG P, REN D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021, 52(8): 8574-8586.

[21] SHEN J, LIU N, SUN H, et al. Vehicle detection in aerial images based on lightweight deep convolutional network and generative adversarial network[J]. IEEE Access, 2019,7: 148119-148130.

[22] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: single shot multibox detector [C]// Computer Vision - ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11 - 14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[23] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. 2023: 7464-7475.