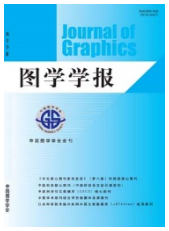

图学学报
Journal of Graphics
ISSN 2095-302X,CN 10-1034/T

《图学学报》网络首发论文

题目：具有双层路由注意力的 YOLOv8 道路场景目标检测方法
作者：魏陈浩，杨睿，刘振丙，蓝如师，孙希延，罗笑南
收稿日期：2023-06-29
网络首发日期：2023-09-25
引用格式：魏陈浩，杨睿，刘振丙，蓝如师，孙希延，罗笑南．具有双层路由注意力的 YOLOv8 道路场景目标检测方法[J/OL]．图学学报．
https://link.cnki.net/urlid/10.1034.T.20230925.1003.002

网络首发时间：2023-09-25 16:49:59
网络首发地址：https://link.cnki.net/urlid/10.1034.T.20230925.1003.002

图 学 学 报
JOURNAL OF GRAPHICS

具有双层路由注意力的 YOLOv8 道路场景目标检测方法

魏陈浩[1]，杨睿[1]，刘振丙[1]，蓝如师[1]，孙希延[2]，罗笑南[2]

(1. 广西图像图形与智能处理重点实验室(桂林电子科技大学) 广西 桂林 541004；
2. 卫星导航定位与位置服务国家地方联合工程研究中心(桂林电子科技大学) 广西 桂林 541004)

摘　　要：随着机动车的数量不断增加，道路交通环境变得更复杂，尤其是光照变化以及复杂背景都会干扰目标检测算法的准确性和精度，同时道路场景下多变形态的目标也会给检测任务造成干扰，针对这一系列问题，提出了一种 YOLOv8n_T 方法，在 YOLOv8 的基础上首先针对骨干网络构建了基于可变形卷积的 D_C2f 块，强化了特征提取网络对复杂背景下目标的特征学习，更好地适应道路目标复杂多变的情形；其次增加了双层路由注意力模块，以查询自适应的方式去除不相关的区域，留下相关度最高的区域；最后针对道路上行人、交通灯等小目标增加小目标检测层，实验表明，提出的 YOLOv8n_T 有效提高了模型在道路场景下的目标检测精度，在 BDD100K 数据集上的平均精度比原始 YOLOv8n 提升了 6.8 个百分点，比 YOLOv5n 提升了 11.2 个百分点。

关　键　词：可变形卷积；道路场景；目标检测；YOLO；注意力机制

中图分类号：TP391　　**文献标识码：**A　　**文 章 编 号：**2095-302X(0000)00-0000-00

YOLOv8 with bi-level routing attention for road scene object detection

WEI Chen-hao[1], YANG Rui[1], LIU Zhen-bing[1], LAN Ru-shi[1], SUN Xi-yan[2], LUO Xiao-nan[2]

(1. Guangxi Key Laboratory of Image and Graphic Intelligent Processing(Guilin University of Electronic Technology), Guilin Guangxi 541004, China;
2. (National Local Joint Engineering Research Center of Satellite Navigation and Location Service(Guilin University of Electronic Technology), Guilin Guangxi, 541004, China)

Abstract: With the continuous increase of motor vehicles, the road traffic environment has become more complex, especially with changes in light conditions and complex backgrounds that can interfere with the accuracy and precision of target detection algorithms. Meanwhile, the diverse shapes of targets in road scenes can also pose challenges to the detection task. In response to this series of issues, a method called YOLOv8n_T has been proposed, which builds on the YOLOv8 skeleton network and incorporates a D_C2f block based on deform-able convolution to enhance feature learning for targets under complex backgrounds, better adapting to the diverse and complex situations of road targets. The model also adds a dual routing attention module to query adaptively and remove irrelevant regions, leaving only the regions with the highest relevance. For small targets such as pedestrians and traffic lights on the road, a small target detection layer is added. Experimental results demonstrate that the proposed YOLOv8n_T significantly enhances the precision of target detection in road scenarios, with an average precision increase of 6.8 percentage points over the original YOLOv8n and 11.2 percentage points over YOLOv5n on the BDD100K dataset.

Keywords: deformable convolution; road scene; object detection; YOLO; attention mechanism

收稿日期：2023-06-29；**定稿日期：**2023-08-17

Received：29 June, 2023；**Finalized**：17 August, 2023

基金项目：国家自然科学基金项目(62172120；62002082)；广西自然科学基金项目(2019GXNSFFA245014；AD20159034)；广西图像图形与智能处理重点实验室项目(GIIP2209)

Foundation items: National Natural Science Foundation of China (62172120； 62002082)； Guangxi Natural Science Foundation (2019GXNSFFA245014； AD20159034)；Guangxi Key Laboratory of Image and Graphic Intelligent Processing Project (GIIP2209)

第一作者：魏陈浩(1999–)，男，硕士研究生。主要研究方向为目标检测，深度学习。E-mail：chwei529@163.com

First author：WEI Chen-hao (1999−), master student. His main research interests cover object detection and deep learning. E-mail：chwei529@163.com

通信作者：蓝如师(1986–)，男，教授，博士。主要研究方向为人工智能、图像处理、医学信息处理。E-mail：rslan2016@163.com

Corresponding author：LAN Ru-shi (1986−), professor, Ph.D. His main research interests cover artificial intelligence, image processing and medical information processing. E-mail：rslan2016@163.com

随着全球经济的发展和城市化进程的加速，机动车的数量不断增加，道路交通环境变得越来越复杂，交通安全问题也日益突出。为了缓解交通安全问题，智能交通系统应运而生。而准确地识别道路上的机动车、行人等目标是智能交通系统中至关重要的任务之一，也是实现交通智能化的基础。

在过去，目标识别算法主要基于传统的机器学习算法，其准确性和性能受到限制。然而，随着深度学习技术的发展和广泛应用，目标识别算法的准确性和性能得到了极大提高。深度学习模型在目标识别任务中表现出色，已经被广泛用于智能交通系统中的目标识别任务中。

随着深度学习技术的兴起，卷积神经网络(CNN)的出现极大地促进了目标检测技术的发展。2012 年，KRIZHEVSKY 等[1]提出了更深层次、参数量更多的 Alexnet 卷积神经网络在 ImageNet 比赛中取得了巨大的成功，随后出现了一系列的 CNN 模型。为了获取更多的特征信息，2014 年 SIMONYAN 和 ZISSERMAN[2]提出了 VGG 网络(Visual Geometry Group)，使用小型池化层和卷积核替换 Alexnet 中的大型池化层和卷积层，并加深网络层数以获取更多的特征信息；网络层数的加深在一定程度上能够更好的拟合特征，同年，SZEGEDY 等[3]利用并行的方法，提出了 GoogLeNet 网络，增加了网络层数，利用大量 1x1 的卷积降低特征图维度，且应用批归一化层对每层数据进行归一化处理，以防止模型过拟合，增强网络的收敛能力； 2015 年，HE 等[4]提出的残差网络(ResNet)解决了随着网络层次加深模型出现性能下降这一问题，该残差网络通过短路连接实现恒等映射，进而解决梯度消失和网络退化的问题。

针对道路场景下人、车辆等小目标的检测算法，主要采用基于深度学习的目标检测算法，其特点是检测速度高的同时平衡了检测精度。2016 年，LIU 等[5]提出了单步多框检测器(Single Shot MultiBox Detector，SSD)，算法借鉴了锚框原理，在特征图上进行预测生成锚框，提高检测精度，该算法虽然保存了高、低层特征图中的目标细节信息和语义信息，但高、低卷积层之间特征信息没有得到很好融和，使得小目标检测过程中提取到的特征信息不全面。

REDMON 等[6]提出 YOLO (You only look once)算法，其思想是直接将输入图像划分成多个单元网格，每个单元网格提取特征获得多个候选框，然后再判断目标的类别、位置和置信度。2017 年 HUANG 等[7]提出了密集网络(Densely Connected Convolutional Networks，DenseNet)其在残差网络的基础上加入了使每一个特征层都能被多次利用的密集连接模块。2018 年，REDMON 和 FARHADI[8]进一步的提出了 YOLOv3 算法，将 YOLOv2 网络中 DarkNet_19 替换成 DarkNet_53 网络结构，采用特征金字塔的网络(Feature Pyramid Network，FPN)结构生成多尺度特征图进行预测，提高了小目标检测效果。为了减少模型参数量和计算损耗，2020 年，BOCHKOVSKIY 等[9]提出了 YOLOv4，采用 CSPDarknet53 特征提取网络，使用 SPP 和特征融合网络，实现了网络的压缩和轻量化。同年，ULTRALYTICS 等提出 YOLOv5，主干网络使用 BottleneckCSP 和聚焦模块进行特征提取，将 YOLOv4 中的 SPP 归到了主干网络，采用特征金字塔网络和路径特征融合相结合的头部网络，模型可以获取到小目标相对更丰富的特征信息。YOLO 系列算法现如今已经发展到 YOLOv8 版本，不仅检测速度快，对大、中型目标的检测效果也优于主流的双阶段目标检测算法。

为了使算法在道路交通场景下具有较高精度、高可移植性和轻量化的特点，本文提出了一种改进 YOLOv8 算法，通过改进特征提取网络，使网络能够提取更加丰富的特征信息，更好的适应道路目标复杂多变的情形。此外，本文加入了双层路由注意力机制，以查询自适应的方式去除不相关的区域，留下相关度最高的区域，在保证性能的同时，具有较高的计算效率。针对复杂的交通场景中目标尺寸较小、难以识别的问题，在网络中添加了小目标检测层，针对更大尺寸的特征图进行局部特征提取，以获得较多的特征细节，同时保持小目标检测的准确率。

1 相关工作

1.1 目标检测

1.1.1 YOLO

YOLO 是一种目标检测算法，于 2015 年提出，与其它目标检测算法不同，YOLO 采用单个神经网络一次性对整个图像进行预测，实现端到端的目标检测。YOLOv1 结构简单，速度快，但不能很好地检测小尺寸的目标。由于 YOLOv1 在小目标检测和物体检测速度上存在问题，YOLOv2 在 YOLOv1 的基础上进行了优化。YOLOv2 使用 Darknet-19 的结构作为基础，并使用了批归一化等方法。此外，YOLOv2 引入了锚框的概念，用于检测不同尺寸的目标，

并使用卷积神经网络的多尺度特征来提高检测精度。YOLOv3 针对 YOLOv2 进行改进，采用了特征金字塔网络提取图像特征，处理早期和晚期的特征图，通过跳跃连接将它们结合起来，从而综合考虑多尺度目标信息，提高检测精度。YOLOv5 于 2020 年 6 月发布，采用的是轻量级特征提取网络，称为 CSPNet-Lite，能够在目标检测方面提供更高的精度和速度。值得一提的是，YOLOv5 还使用强化学习算法对模型进行自适应训练，从而使模型更加适应各种目标检测任务。虽然 YOLOv5 在目标检测方面取得了很大的进展，但在目标密集时容易出现漏检和误检的情况，对于具有复杂形状和纹理的目标及小物体的特征提取能力较弱，导致物体被忽略，检测精度较低。

1.1.2 SSD

SSD 是一种基于深度学习的目标检测算法，与 YOLO 算法类似，SSD 也是一种端到端的目标检测算法，可以同时进行目标类别识别和边界框检测。SSD 使用一个卷积神经网络对图像进行特征提取，并在多个特征图上进行边界框检测。通过使用不同大小的卷积核，它可以检测不同尺度的目标。SSD 使用多个大小和比例的默认框来覆盖整个输入图像，因此可以在不需要额外的区域建议网络[10] (Region Proposal Network，RPN)的情况下检测目标。此外 SSD 还使用数据增强技术来增加数据的多样性，包括亮度、对比度、变形和裁剪等操作。这些操作可以增加训练数据的数量，从而提高检测准确率。总的来说，SSD 算法是一种快速，准确，具有端对端检测能力的目标检测算法，其设计思想以及特征提取，多尺度检测和非极大值抑制等技术为目标检测的精度和速度提供了保障。同时 SSD 算法也存在处理小目标时精度较低的问题。

1.1.3 Faster-RCNN

Faster-RCNN 是一种基于 RPN 的快速目标检测算法。相对于以往的 R-CNN、Fast-RCNN，Faster-RCNN 具有更快的速度和更高的准确率。Faster-RCNN 通过引入区域建议网络来共享 卷积计算并快速生成 RoI(区域兴趣)提议，从而大大减少了每张图像的区域提议计算量。Faster-RCNN 由于需要两个不同的网络进行训练，因此训练时间较长，此外网络需要在多个候选框上运行非极大值抑制以去除多余的框，导致检测速度较慢。

1.2 注意力机制

计算机视觉中注意力机制一般是根据输入图像的特征进行动态权重调整的处理。

在深度神经网络中，不同特征图中的不同通道通常表示不同的对象。HU 等[11]首先提出了通道注意力的概念，并为此目的提出了 SENet，网络的核心是一个挤压和激励块，用于收集全局信息，捕获通道关系，并提高表示能力。由于卷积神经网络通常具有巨大的计算成本，特别是对于较大的图像输入。为了将有限的计算资源集中在重要区域上，JADERBERG 等[12]提出了空间注意力网络(Spatial Transformer Networks，STN)，其引入了一个新的可学习的空间转换模块，在无需修改和额外的训练监督或优化过程，就能够插入到现有的卷积结构中，使特征图能够进行空间变换。通道注意力和空间注意力可以通过串联或并联的方式进行组合成为混合注意力机制，WOO 等[13]提出了一种轻量级的注意力模块 CBAM，可以在通道和空间维度进行注意力操作。

在目标检测领域，注意力机制可以帮助模型处理更加复杂的场景和任务，提高模型的鲁棒性和可解释性，在目标检测任务中应用注意力机制可以让模型能够关注目标周围的特征，还可以帮助模型捕获和分离不同身份之间的特征，从而使得检测结果更加精确。尽管注意力机制在目标检测任务中效果较好，但由于需要对不同位置的输入进行计算和加权，导致计算开销较大。

2 本文方法

图 1 展示了本文所提出的 YOLOv8n_T 方法的网络结构图，本网络由 4 个部分构成，分别是输入，主干，头部和输出。在输入端 YOLOv8 通过对输入图像进行马赛克数据增强，自适应图像缩放、拼接、叠加等数据增强方法来提高模型的泛化能力。

本文基于可变形卷积构建了 D_C2f 模块来替换骨干网络中的 C2f 模块，以提高模型对不规则物体的识别能力，同时在骨干网络尾部增加了双层路由注意力模块以进一步提升模型的关注度和泛化性能。

头部网络采用了特征金字塔网络和路径聚合网络的网络结构以增强特征融合能力，本文在头部增加了小目标检测层(图中虚线部分)，以进一步提高模型对小目标识别的准确率。

本章内容依次针对 D_C2f、双层路由注意力和小目标检测层进行描述。

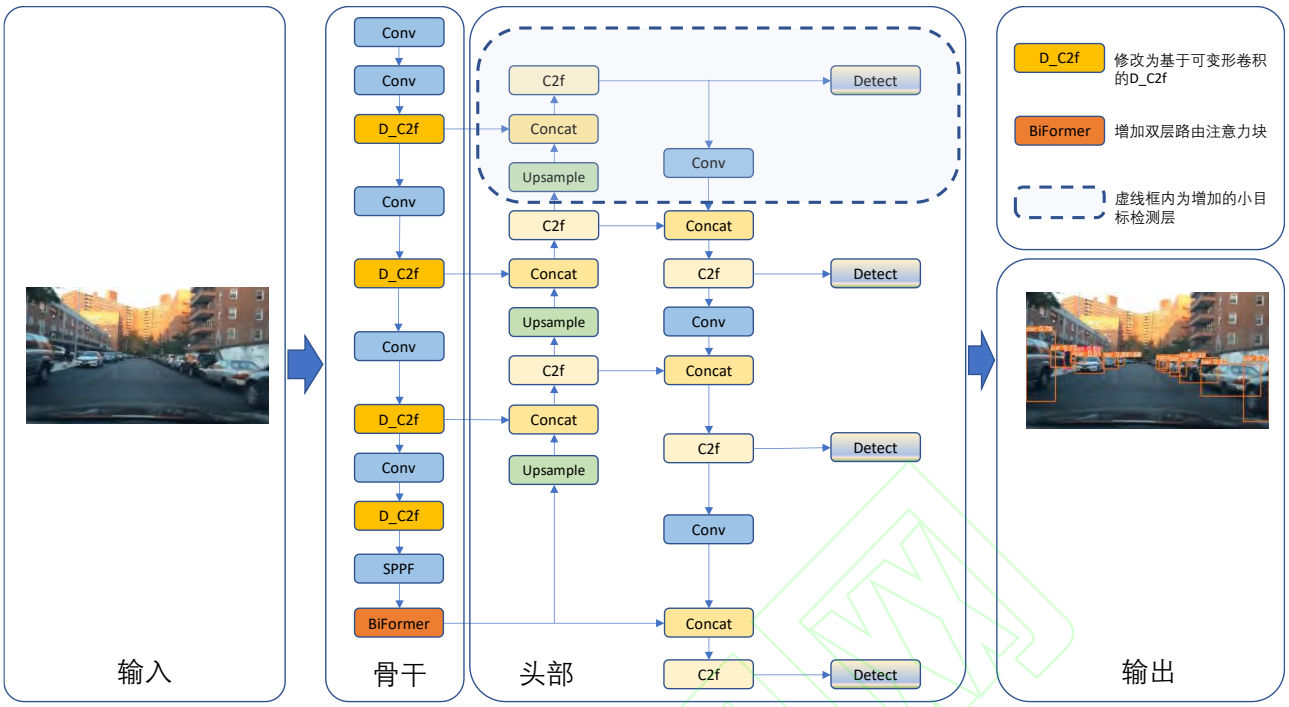

图 1　YOLOv8n_T 网络结构图

Fig. 1　YOLOv8n_T network architecture

2.1　D_C2f

在道路场景下，车辆行人等目标多样且复杂，由于目标的大小、形状、位置、方向等都育有一定的变化性，因此使用传统的卷积操作往往难以准确地捕获目标的准确位置，甚至可能会导致目标漏检或误检的问题。为了解决上述问题，本文采用 DAI 等 [14] 提出的可变形卷积网络 (Deformable Convolution)，重新构建了 YOLOv8 算法中 C2f 网络结构来提升网络的检测能力。

在传统卷积中，每个卷积核都是固定形状的，因此无法处理物体形变的情况。而可变形卷积中，每个卷积核不再是一个固定的矩形，而是由一个基础网格和一组偏移量共同组成的可变形矩形。在进行卷积操作时，可以根据形状偏移量动态地调整卷积核的形态，从而更好地适应物体的形变。

假设输入特征图为 $I \in R^{C_{in} \times H \times W}$，对输入特征图中每个在 p_0 位置的标准卷积可写作如下形式

$$y(p_0) = \sum_{p_n \in R} \omega(p_n) x(p_0 + p_n) \tag{1}$$

其中，p_n 为 p_0 点在卷积核范围内的偏置项，是一个常数，用于帮助模型学习偏差，提高模型的准确性和稳定性。可变形卷积在上述标准卷积的基础上为每一个位置都引入了一个偏移量 Δ_{p_n}，并通过下式进行计算

$$y(p_0) = \sum_{p_n \in R} \omega(p_n) x(p_0 + p_n + \Delta_{p_n}) \tag{2}$$

其中，Δ_{p_n} 由输入特征图经过一个卷积层生成，通常为小数，ω 为采样点权重。

由于加入偏移量后的位置往往会出现小数，而这些位置并不对应于特征图上实际存在的像素点，因此需要使用插值的方法来得到偏移后的像素值。为解决上述问题，通常使用双线性插值来实现，即在特征图上找到四个最近的像素点，接着采用加权平均的方式计算偏移后的像素值。

$$I_{x',y'} = \sum_{i=1}^{2} \sum_{j=1}^{2} \omega_{i,j} \cdot I_{x+i-1, y+j-1} \tag{3}$$

其中，I 表示输入特征图像素，$I_{x',y'}$ 表示计算后的偏移像素值。

图 2 展示了根据可变形卷积构建的 Bottleneck，由两层可变形卷积层顺序连接，最后叠加输入特征图得到输出结果。图 3 是使用 D_Bottleneck 重建的 D_C2f 模块，由卷积层、分离层和 D_Bottleneck 组成。

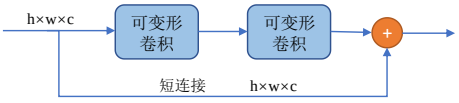

图 2　基于可变形卷积的 D_Bottleneck

Fig. 2　D_Bottleneck based on deformable convolution

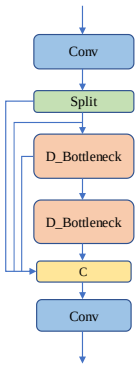

图 3　基于 D_Bottleneck 的 D_C2f 模块

Fig. 3　D_C2f module based on D_Bottleneck

2.2　双层路由注意力

MHSA(多头自注意力机制)是视觉 Transformer 中实现注意力机制的关键部分之一，MHSA 可通过增加头数来提高模型的表现。然而，增加头数会带来一些可伸缩性问题，同时模型的计算和内存需求也会增加。这使得多头注意力机制在更大的图像输入上变得极其耗时，并且需要大量的计算资源。上述问题同时限制了视觉 Transformer 的扩展能力，难以进行更大规模的计算机视觉任务。

双层路由注意力[15]是一种动态的，查询感知稀疏注意力机制，主要思想是在粗级别区域筛选出大部分不相关的键值对，只留下少量的路由区域，随后在这些路由区域应用令牌到令牌注意力得到最有相关性的区域。如图 4 描述了双层路由注意力机制通过在前 k 个相关窗口收集键值对计算最相关区域。假设给定一个二维输入特征映射，首先将其划分为 $S\times S$ 个非重叠区域，然后使用线性投影将导出为原始 $\boldsymbol{QKV}$ 张量，即

$$Q = X^r W^q \tag{4}$$

$$K = X^r W^k \tag{5}$$

$$V = X^r W^v \tag{6}$$

其中，W^q，W^k，W^v 分别是 Q，K，V 投影权重，X^r 表示给定特征图的子区域。

为了确定每个给定子区域应该关注哪些区域，首先计算 Q 和 K 的单区域平均值获得区域级的查询和关键字 Q^r，K^r；然后将 Q^r 和 K^r 转置矩阵进行矩阵相乘获得区域到区域的邻接矩阵 A^r，邻接矩阵 A^r 表示两个区域的关联程度，因此通过 TopK 算法保留每个区域中前 k 个连接即可去除关联程度较弱的部分

$$A^r = Q^r (K^r)^T \tag{7}$$

$$I^r = topK(A^r) \tag{8}$$

其中，I^r 表示路由索引矩阵，包含了前 k 个最相关的连接，由于 k 个区域是分散在整个特征图上，为了高效的处理首先收集键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 张量，即

$$K^g = gather(K, I^r) \tag{9}$$

$$V^g = gather(V, I^r) \tag{10}$$

其中，$gather(\cdot,\cdot)$ 函数根据索引矩阵从输入张量中提取数据并生成新的组合后的张量。

最后将 K^g，V^g 应用注意力机制，得到经过注意力的特征图输出 O，并写作

$$O = Attention(Q, K^g, V^g) \tag{11}$$

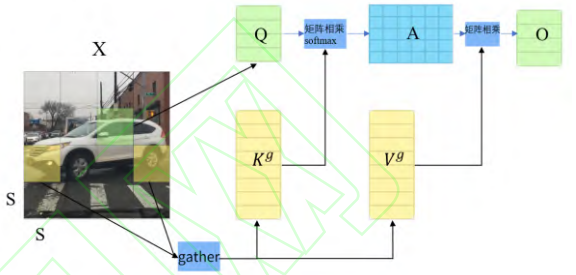

图 4　双层路由注意力机制

Fig. 4　Bi-level routing attention mechanism

2.3　小目标检测层

道路场景通常包含许多小目标，如车牌、交通信号灯、行人等。YOLOv8 进行特征融合后检测层输出尺寸不同的特征图用于检测不同尺寸的目标。尺寸大的特征图感受野较小，蕴含的目标位置和局部特征细节较多，适合检测小目标；尺度小的特征图感受野较大，语义信息丰富但局部细节不明显，适合检测大目标。原始 YOLOv8 中最大特征图仅为 80×80，为了提高道路小目标检测的精度，同时避免下采样导致的细节特征丢失过多，本方法针对原始 YOLOv8 网络结构进行改进，由图 1 所示，虚线框部分为增加的小目标检测层，在原网络主干的 P2 层设置特征图流出，增加一层特征图尺寸为 160x160 的小目标检测层，在降低了小目标的漏检和误检率条件下，进行更深层次的特征传递和特征融合，使得小目标定位与识别更为准确。

3　实验结果与分析

本章通过实验评估了所提出的改进 YOLOv8 算法在道路场景下目标检测任务的有效性。具体来说，本方法在 BDD100K 和 NEXET 数据集上对改进 YOLOv8 及六种主流目标检测方法进行训练和测试，以对比每个目标检测算法的性能。同时还对每个改进算法(分别是在原始 YOLOv8n 单独替换

了 D_C2f 模块的 YOLOv8n_D,单独添加了双层路由注意力模块的 YOLOv8n_BRA，单独增加小目标检测层的 YOLOv8n_nano 和本文提出的同时添加了上述创新的 YOLOv8n_T)进行消融实验。同时也选取了 BDD100K 数据集中不同场景下的图片，比较了多种目标检测算法在实际场景中的检测效果。对每一种算法统一设置 epochs 为 100。

3.1 数据集与实验参数设置

BDD100K 数据集[16]是由加州大学伯克利分校发布的用于自动驾驶领域研究的大规模、多样化的数据集。数据集中图像与视频从全球不同地区的城市和公路收集而来，涵盖了各种不同的场景，包括城市街道、高速公路、乡村小路、人行道等 10 类物体，同时数据集的标注信息非常丰富，有助于进行更加精确的目标检测和跟踪。为了提高数据集的多样性，每张图片进行了多次标注，每种方式都包含了不同的角度、尺度、变换和遮挡，以模拟真实世界中的视觉场景。我们从 70000 张图片中随机选择了同时包含至少两类目标的 22311 张图片进行网络训练，其中 18071 张为训练集，2008 张为验证集，2232 张为测试集。

NEXET[17]是一个交通场景下车辆检测数据集，包含了 500000 张用于车辆检测的图像，类别数量为 5，同样筛选了目标数量大于 2，目标类别大于 2 的 20000 张图片进行训练，其中 16200 张为训练集，1800 张为验证集，2000 张为测试集。

实验采用 Ubuntu18 操作系统，CPU 为 Intel Xeon Silver 4110，GPU 为 Nvidia RTX2080ti，12G 显存；深度学习框架为 pytorch 1.13.1，CUDA 11.1。

3.2 评价指标

mAP 是一种用于衡量目标检测算法性能的常用指标，它通过计算检测算法在不同置信度阈值下的平均精度(Average Precision)来综合评估算法表现。本文使用 mAP50， mAP50-95，APs，APm，APl 作为主要的评价指标。

在目标检测中，一般使用交并比指标来评估预测框和真实标注框之间的重叠程度。在进行 mAP 计算时，需要将所有检测结果按照置信度从高到低排序，并依次计算每个预测框在不同置信度阈值下的精度和召回率。然后将所有不同置信度阈值下的精度-召回率曲线(Precision-Recall Curve)下的面积(Area under Curve，AUC)求和并除以类别数，即可得到 mAP 值，即

$$mAP = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{1} P(R)dR \tag{12}$$

其中，P 是指确切检测到目标的预测框占所有预测框中的比例，R 是指实际检测到目标的预测框占所有真实标注框中的比例。因此，P-R 曲线展示了准确率和召回率在不同阈值下的变化情况，而 AUC 则是曲线下方的面积，用于度量算法在不同置信度阈值下的总体表现，即

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

其中，TP 表示模型预测为正样本且与真实标注匹配的检测框数量；FP 表示模型预测为正样本但与真实标注不匹配的检测框数量；FN 表示模型未能检测到的真实标注数量。

3.3 对比实验

本文在目标检测对比实验中使用了七种经典的目标检测方法，包括 SSD，Faster-RCNN，YOLOv8n，YOLOv5n，YOLOv7-Tiny，FCOS[18]和本文所提出的 YOLOv8n_T，并通过 mAP50，mAP50-95，APs，APm，APl 五个指标来评估它们的性能。实验结果见表 1，可以看出本文所提出的改进 YOLOv8n_T 在所有指标上均表现出最佳性能，mAP50 指标下比次优方法 Faster-RCNN 高了 4 个百分点，同时在 APs 指标上明显优于其他方法。FCOS 在两个数据集上的平均精度比 Faster-RCNN 稍差，但在 APs 指标上优于 Faster-RCNN，原因可能是 FCOS 采用了完全卷积的网络结构同时直接预测像素点的物体中心。其中，YOLOv5 表现最差，可能存在的原因是由于其采用了较浅的特征提取网络，因此平均精度结果值均为最低。

表 1 YOLOv8n_T 与其他目标检测算法的实验结果对比

Table 1 Comparative experimental results between YOLOv8n_T and other object detection algorithms

方法	BDD100K					NEXET				
	mAP50	mAP50-95	APs	APm	APl	mAP50	mAP50-95	APs	APm	APl
YOLOv5n	0.51	0.284	0.095	0.279	0.384	0.628	0.435	0.106	0.365	0.498
SSD	0.528	0.297	0.098	0.283	0.379	0.632	0.462	0.11	0.382	0.505
YOLOv7-Tiny	0.547	0.316	0.112	0.298	0.406	0.643	0.453	0.125	0.397	0.516

YOLOv8n	0.554	0.347	0.115	0.294	0.415	0.651	0.474	0.133	0.421	0.528
FCOS	0.564	0.33	0.127	0.279	0.397	0.647	0.461	0.162	0.416	0.524
Faster-RCNN	0.587	0.362	0.106	0.316	0.42	0.66	0.482	0.118	0.408	0.531
YOLOv8n_T	**0.622**	**0.401**	**0.193**	**0.371**	**0.458**	**0.684**	**0.496**	**0.251**	**0.432**	**0.560**

注：加粗数据为最优值

3.4 消融实验

为了验证算法改进的有效性，本节设计了 YOLOv8n ， YOLOv8n_D ， YOLOv8n_BRA ， YOLOv8n_nano 以及 YOLOv8n_T 的消融实验，其中 YOLOv8n_D 是在 YOLOv8n 的骨干网络中使用基于可变形卷积的 D_C2f 块替换普通的 C2f。YOLOv8n_BRA 是通过在 YOLOv8n 的骨干网络尾部添加了双层路由注意力块实现。YOLOv8n_nano 模型则是仅在原始 YOLOv8n 的骨干网络 P2 层增加特征图分流，添加针对小目标的检测层来实现。YOLOv8n_T 是本文所提出的算法，包含了 D_C2f，双层路由注意力和小目标检测层的改进。

消融实验对比结果见表 2，针对道路复杂场景下的目标检测，改进算法的每个阶段性能均有提升，且效果明显。修改了基于可变形卷积的 D_C2f 块的 YOLOv8n_D 模型比 YOLOv8n 在 mAP50 和 mAP50-95 上分别提升了 2 个百分点和 1 个百分点，表明基于可变形卷积的 D_C2f 模块在进行特征提取时能够更有效的获得丰富的梯度流信息。添加双层路由注意力机制的 YOLOv8n_BRA 模型在 mAP 上相对原始 YOLOv8n 也有所提升，可以看出双层路由注意力机制的添加有效的改善了在处理不同尺度目标以及模糊背景时特征难以提取的问题。增加了小目标检测层的 YOLOv8n_nano 模型在 mAP50 上有 5 个百分点的提升，表明小目标检测层的添加在特征图较大时进行分流，避免了在特征提取过程中小目标实例信息丢失的问题，提高了小目标检测的准确率。

表 2 消融实验结果

Table 2 Experiment results for each component

方法	D_C2f	BRA	nano	mAP50	mAP50-95
YOLOv8n	×	×	×	0.554	0.347
YOLOv8n_D	√	×	×	0.571	0.356
YOLOv8n_BRA	×	√	×	0.557	0.35
YOLOv8n_nano	×	×	√	0.608	0.393
YOLOv8n_T	√	√	√	**0.622**	**0.401**

注：加粗数据为最优值

3.5 实验分析

SSD， YOLOv5n， YOLOv7-Tiny， FCOS 和 YOLOv8n_T 五种算法在 BDD100K 数据集上的处理结果如图 5 所示，从上至下依次为图一至图四。如图所示，在每个算法都能准确的识别较大的目标，如近处的行人，车辆等。在图三的夜晚光照强度较低时，SSD，YOLOv5n 和 YOLOv7-Tiny 都未能识别到图片左侧的模糊人像，而 FCOS 和 YOLOv8n_T 都能准确的识别。SSD 和 YOLOv5 表现接近且性能较差，图一左上角交通标志目标较小，未准确的识别到，同时在图 4 中 SSD 相对 YOLOv5 漏检较多。相对于其他集中方法，本文 YOLOv8n_T 在误检率上也有效提升，FCOS 将图一右侧的顶棚错误的识别为交通标志，而其余 3 种方法都未检测这一错误目标。因此本文所提出的 YOLOv8n_T 性能明显优于其他 3 种方法，上述所提到的难以检测的目标均被正确的检测。

(a) (b) (c) (d) (e)

图 5 复杂道路场景实验结果对比

Fig. 5 comparative experimental results of complex road scene ((a) SSD; (b) YOLOv5n; (c) YOLOv7-Tiny; (d) FCOS; (e) YOLOv8n_T)

3.6 数据分析

为了比较不同改进对不同类型物体的检测效果，本节记录了在 BDD100K 数据集中 10 类目标的 mAP，结果如图 6 所示，由于 train 目标在本文数据集中数量较少，可能会导致偏差过大，因此忽略此项。实验结果表明在单独添加双层路由注意力和小目标检测层后，对交通信号灯这类小目标识别精度有明显提升。改进后的算法对汽车、卡车等较大物体的识别效果稳定，对交通灯、行人和骑手等较小目标识别效果呈稳定上升的走势。

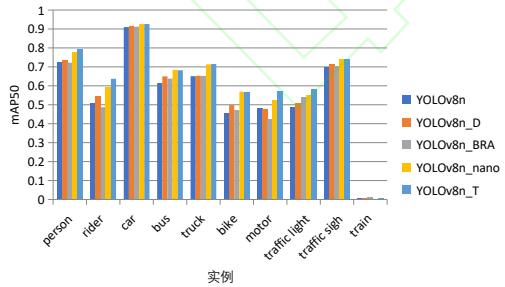

图 6 不同目标上的 mAP 比较结果

Fig. 6 comparative results of mean average precision (mAP) on different targets

4 结论

本文提出了一种改进 YOLOv8 算法用于道路场景下的目标检测，该算法的主要贡献在于，通过对 YOLOv8 网络的改进，采用能够提取更加丰富的特征信息的可变形卷积构建 C2f 块，更好的适应了道路目标复杂多变的情形。同时，双层路由注意力机制被引入，以自适应地去除不相关的区域，从而保留相关度最高的区域，并在保持性能的同时具有较高的计算效率。此外，针对复杂交通场景中大小较小、难以识别的小目标，本文通过添加小目标检测层，对较大尺寸的特征图进行局部特征提取，能够获得更多的特征细节，从而大大提高了小目标检测的准确率。在 BDD100K 和 NEXET 数据集上进行了测试和比较，通过实验分析，证明了各部分改进的可行性，YOLOv8n_T 在平均精度和准确率等指标方面都由于其他目标检测方法，在复杂场景下小目标也能够精准的检测。

尽管本文算法精度较高，但准确率仍有很大的改进空间，同时算法部署到算力有限的边缘设备中仍有不小的挑战，今后的工作中将继续研究如何在轻量化算法的同时提高算法的检测精度。

参考文献 (References)

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2023-05-10]. https://arxiv.org/abs/1409. 1556.

[3] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with

convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE Press, 2015: 1-9.

[4] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 770-778.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Computer Vision–ECCV 2016. Amsterdam: Springer International Publishing, 2016: 21-37.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 779-788.

[7] HUANG G, LIU Z, VDM L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017: 4700-4708.

[8] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. [2023-05-10]. https://arxiv.org/abs/1804.02767.

[9] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. [2023-05-10]. https://arxiv.org/abs/2004.10934.

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

[11] 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.

[12] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[J]. Advances in Neural Information Processing Systems, 2015, 28.

[13] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[14] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision, Venice: IEEE Press, 2017: 764-773.

[15] ZHU L, WANG X, KE Z, et al. BiFormer: Vision Transformer with Bi-Level Routing Attention[C]// 2023 IEEE Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE Press, 2023: 10323-10333.

[16] YU F, CHEN H, WANG X, et al. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning [C]// 2020 IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Press, 2020: 2636-2645.

[17] KLEIN I. NEXET-The Largest and Most Diverse Road Dataset in the World[EB/OL]. [2023-05-10]. https://www.kaggle.com/datasets/solesensei/nexet-original

[18] TIAN Z, SHEN C, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE Press, 2019: 9627-9636.