唐坤俊, 宁媛, 刘聂天和. 基于改进 RT-DETR 的道路交通场景检测[J]. 智能计算机与应用,2025,15(9):82-89. DOI:10.20169/j.issn.2095-2163.250913

基于改进 RT-DETR 的道路交通场景检测

唐坤俊[1], 宁 媛[1], 刘聂天和[2]
(1 贵州大学 电气工程学院, 贵阳 550025; 2 贵州电网有限责任公司贵阳花溪供电局, 贵阳 550025)

摘 要: 由于道路交通场景存在许多遮挡和小目标物体,很容易出现误检测和漏检,因此提出一种基于 RT-DETR 的改进目标检测模型来提升检测性能。在特征提取网络方面,采用经过 Shuffle Attention(SA)注意力机制增强的 ResNet-18,用来加强网络特征提取能力;同时引入 Cascaded Group Attention(CGA)机制替换原模型尺度内特征交互(AIFI)模块中的多头自注意力机制(MHSA),成功减少计算冗余,提升了模型性能。最后,构建专门针对道路交通场景的数据集进行实验。模型在 RTX4070ti GPU 平台上进行了性能验证。性能评估表明,改进后的模型在平均精度(mAP)上达到 72.9%,较原 RT-DETR 模型提升 2.1%。此外,在每秒帧数(FPS)方面,改进模型同样表现出色、达到 132.1,优于 RT-DETR 9 帧和 YOLOv8m 23 帧。综合实验结果显示,本研究提出的改进模型不仅保持了高检测精度,还成功地加速了模型计算。这些改进对于实时且精确处理道路交通场景的目标检测具有重要的实用价值。

关键词: 目标检测; 深度学习; RT-DETR; 道路交通场景检测; 注意力机制
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2025)09-0082-08

Road traffic scene detection based on the improved RT-DETR

TANG Kunjun[1], NING Yuan[1], LIUNIE Tianhe[2]
(1 School of Electrical Engineering, Guizhou University, Guiyang 550025, China;
2 Guiyang Huaxi Power Supply Bureau of Guizhou Power Grid Co., Ltd., Guiyang 550025, China)

Abstract: Due to the prevalence of occlusions and small targets in road traffic scenarios, there is a high propensity for false detections and omissions. Consequently, the paper proposes an enhanced object detection model based on RT-DETR. In the realm of feature extraction, the proposed model employs a ResNet-18 framework augmented with a Shuffle Attention (SA) mechanism, bolstering its feature extraction capacity. Additionally, the research integrates a Cascaded Group Attention (CGA) mechanism, substituting the Multi-Head Self-Attention (MHSA) within the original model's Attention Intra-Feature Interactions (AIFI) module, thereby significantly reducing computational redundancy and enhancing model performance. Experiments are conducted using a dataset specifically designed for road traffic scenes. The model's performance is validated on an RTX4070ti GPU platform. Performance evaluation reveals that the improved model achieves a mean Average Precision (mAP) of 72.9%, marking a 2.1% increase compared to the original RT-DETR model. Moreover, in terms of Frames Per Second (FPS), the enhanced model reaches 132.1 FPS, surpassing RT-DETR by 9 frames and YOLOv8m by 23 frames. The comprehensive experimental results demonstrate that the proposed improved model not only maintains high detection accuracy, but also significantly accelerates model computation. These advancements hold significant practical value for real-time and precise object detection in road traffic scenarios, leveraging the capabilities of deep learning in object detection.

Key words: object detection; deep learning; RT-DETR; road traffic scene detection; attention mechanism

0 引 言

随着经济发展和社会进步,自动驾驶和智慧交通技术的研究与发展受到广泛的关注。实时且准确的目标检测是自动驾驶和智慧交通系统的重要组成部分,但是实际道路情况是复杂多变的,对目标检测算法提出了更高的挑战。

近年来,目标检测作为计算机视觉的基础任务一

基金项目: 贵州省科技计划基金(黔科合 ZK2022135)。
作者简介: 唐坤俊(1998—),男,硕士研究生,主要研究方向:深度学习目标检测,模型轻量化部署; 刘聂天和(1998—),男,硕士研究生,主要研究方向:电网技术,边缘计算。
通信作者: 宁 媛(1968—),女,教授,硕士生导师,主要研究方向:计算机视觉,图像处理。Email:1317045666@qq.com。
收稿日期: 2023-12-17

直备受业界瞩目。在深度学习领域，以卷积神经网络（CNN）和 Transformer[1] 为基础的方法自问世以来就引发关注，已经成为目标检测领域的主流技术路径。基于 CNN 的经典检测器方案是使用卷积神经网络作为特征提取 Backbone，然后使用手工组件 Anchor-Base（Fast R-CNN[2]、Faster R-CNN[3]、YOLOv1～v7[4-8] 等）或者 Anchor-Free（YOLOv8，FCOS[9]，YOLOX[10] 等）加上非极大值抑制（NMS）来筛选最终的候选框。然而 Anchor-Base 或 Anchor-Free 的 2 种方案都利用非最大抑制进行后处理，这给经典检测器带来了推理性能的瓶颈。此外，由于非极大值抑制不使用图像信息，因此在边界框保留和删除中容易出错。近年来，Transformer 已广泛应用到计算机视觉的物体分类领域，例如 Vision Transformer[11]、Swin Transformer[12] 等。Transformer 用在目标检测领域的开山之作：DETR[13]（DEtection TRansformer），消除了传统检测流程中的 Anchor 和非极大值抑制（NMS）组件，通过二分匹配直接预测检测对象，简化了检测流程。尽管 DETR 具有显著优势，但其存在训练收敛慢和查询优化难的问题。为解决这些问题，出现了多个变体，如 Deformable-DETR[14] 通过提高注意力机制的效率来加速训练收敛，DAB-DETR[15] 引入了 4D 参考点以优化预测框，而 DINO[16] 在此基础上取得了较为先进的成果；RT-DETR[17] 则是解决了标准 DETR 模型的高计算成本问题。无论是基于 CNN、还是基于 Transformer 的模型，都在不断演化以应对日益复杂和动态的视觉场景。然而，道路交通场景还存在许多密集遮挡目标和小目标导致的误检、漏检问题[18]，影响模型检测性能。

为了能够将现有更高效的目标检测模型应用在道路交通场景，常用的改进方法有：使用更强的模型特征提取网络、添加小目标检测头、引入注意力机制、改变特征融合方式等。盛博莹等学者[19] 以 YOLOv5s 为基础框架，使用反馈机制的特征提取网络 RFP-PAN，以提高小目标检测精度。冉险生等学者[20] 通过改进网络特征融合的方式提升了检测过程中密集遮挡目标、小尺度目标出现的漏检和误检问题，但是模型结构复杂，不利于实时场景。李轩等学者[21] 为了解决密集目标的遮挡问题，提出了 Occlusion Loss，通过提高预测框和真实框的匹配程度以使定位更加准确。李永上等学者[22] 改进回归损失函数以加快边界框回归速率，并改进非极大抑制，改善小目标的漏检问题。以上的改进对传统目标检测 CNN 模型进行了改进，效果十分显著和有效。综合卷积神经网络的优点，将其运用在端到端的模型，同时加上 DETR 类模型不需要 NMS 组件的优势，不容易在边界框保留和删除中出错，所以本文对现有 RT-DETR 模型进行了关键性改进，强化其在处理实时道路交通场景时的性能。改进如下：

（1）在特征提取的骨干网络 ResNet-18 中，加入 Shuffle Attention 机制，显著增强了模型对关键特征的提取能力，提升小目标和遮挡目标识别的准确性。

（2）通过引入 Cascaded Group Attention（CGA）机制，替换模型的尺度内特征交互 AIFI 模块中（Encoder）的多头自注意力机制，使 MHSA 能够更加高效和快速地运行，降低了计算冗余。

1　RT-DETR 介绍

RT-DETR 由 Lv 等学者[17] 于 2023 年 4 月首次提出，是一种较新的实时目标检测模型，主要包括：特征提取骨干网络（Backbone）；高效混合编码器（Hybrid Encoder）和解码器（Decoder）。RT-DE 原始模型结构如图 1 所示。

骨干网络（Backbone）在检测模型中担当着特征提取的重要角色，在本次实验中选择了基于 CNN 的 ResNet-18 模型作为核心。这一选择考虑到 ResNet-18 既具有较低的参数量，又拥有强大的特征提取能力，从而有效平衡了后续 Transformer 编码器-解码器的参数。骨干网络的最后 3 个阶段（S3、S4、S5）的输出特征，被用作后续混合编码器（Hybrid Encoder）的输入。

在高效混合编码器（Hybrid Encoder）部分，Lv 等学者[17] 通过分析 Transformer Encoder 的计算量，解耦了基于 Transformer 的全局特征编码，从而设计了尺度内特征交互（AIFI）和跨尺度特征融合模块（CCFM）结合的新的高效混合编码器（Efficient Hybrid Encoder），这在一定程度上类似于经典目标检测模型中常用的特征金字塔网络（FPN）。此外，为了进一步优化计算效率，把 Encoder Layer 的层数从 6 减小到 1 层。这样的设计使得模型能更精准地理解图像中的语义内容，增强多个尺度上的目标定位能力，从而显著提高目标检测的准确性。

解码器部分使用 IoU 感知查询选择机制，这种机制从 Encoder 的输出中选出固定数量的 Token 作为 Decoder 的初始查询。在解码阶段，这些查询经过连续迭代优化，有效生成最终的边界框和置信度分数。不同于传统目标检测模型，这里无需通过 NMS 进行候选框筛选，实现了真正的端到端目标检测。

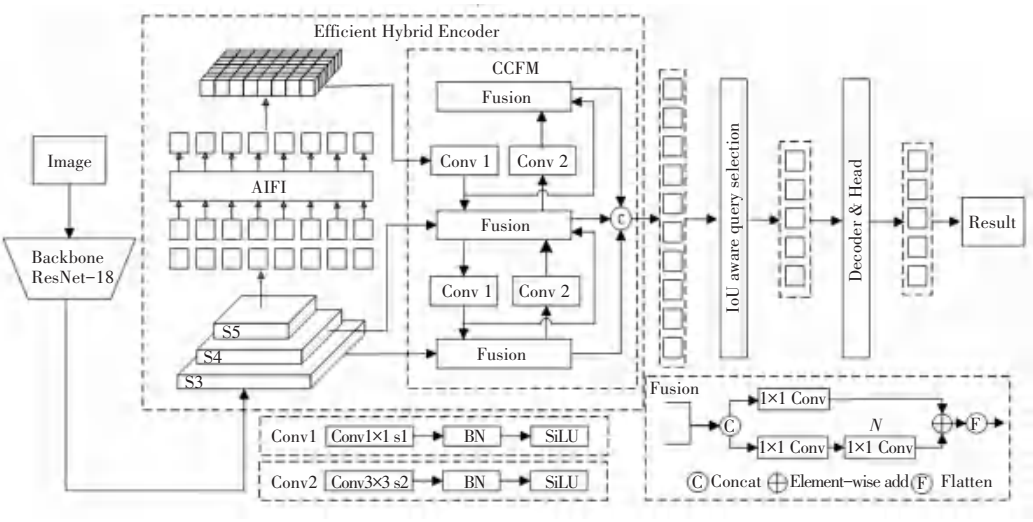

图 1 RT-DETR 原始模型

Fig. 1 Original model of RT-DETR

2 RT-DETR 改进策略

本文的核心改进集中在采用先进的注意力机制，旨在提升模型的效率和精确度。首先，在 Backbone 部分，选择了 ResNet-18[23-25] 网络，因为其具有较低的模型参数量，有助于整体模型的轻量化。特别地，在基础模块中加入了 Shuffle Attention[15] 机制，该机制通过优化特征提取过程，进一步提升了模型对小目标和遮挡目标的识别能力。其次，从 EfficientViT[16] 模型中借鉴了级联分组注意力（Cascaded Group Attention，CGA）模块。这一模块的创新之处在于为每个注意力头提供不同的输入，有效减少了多头自注意力（MHSA）机制中的计算冗余。这种设计不仅提高了计算效率，还增强了模型在处理复杂场景时的准确性和实时性。

2.1 Backbone 的改进

ResNet-18 是一种在深度学习领域广泛应用的网络架构，属于残差网络（ResNet）系列。由微软研究院的 He 等学者[25] 于 2016 年提出，这一架构的设计初衷是为了解决深度神经网络中的梯度消失问题。ResNet-18 因其较浅的网络深度和较少的参数而成为焦点，使其成为计算资源有限环境下的理想选择。残差模块结构如图 2 所示。

Shuffle Attention[23] 是一种专门针对计算机视觉设计的注意力机制，主要集中于图像的通道级别特征。SA 注意力设计如图 3 所示。这种机制通过学习不同通道之间的相互关系，有效地选择并强调图像特征中的关键信息。通过引入通道重排加强了不同特征之间的交互，从而增强了整体网络的表达能力。此外，还允许模型自适应地学习和调整通道权重，使其能更好地适应不同的任务和输入情境。这样的设计有效减少了冗余梯度信息的干扰，提高了模型在目标检测等任务中的性能表现。

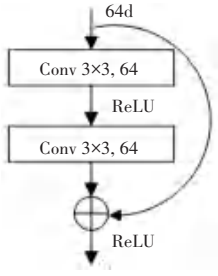

图 2 残差模块

Fig. 2 Residual module

Shuffle Attention 的计算过程如下：

首先将输入特征图 X 分为 G 组，每组有 c/g 个通道，其中 c 是通道数。对此可以表示为：

$$X = [X_1, X_2, X_3, \cdots, X_G] \tag{1}$$

对第 i 个 c/g 特征图拆分为 2 部分：X_1 和 X_2，每部分有 $c/2g$ 个通道，对 X_1 使用通道注意力，X_2 使用空间注意力机制。由此得到：

$$\begin{aligned} &\mathrm{Fca}(X1) = \sigma(\mathrm{FC}(\mathrm{AvgPool}(X1))) \\ &\mathrm{Fsa}(X2) = \sigma(\mathrm{Conv}(\mathrm{Concat}(\mathrm{AvgPool}(X2), \mathrm{MaxPool}(X2)))) \end{aligned} \tag{2}$$

其中，σ 表示 Sigmoid 函数；FC 表示全连接层；AvgPool 表示全局平均池化；Conv 表示卷积层；Concat 表示连接操作。

接下来，将注意力权重应用到相应的通道上，由此推得：

$$X_1' = \mathrm{Fca}(X_1) \cdot X_1 \tag{3}$$

$$X_2' = \mathrm{Fsa}(X_2) \cdot X_2 \tag{4}$$

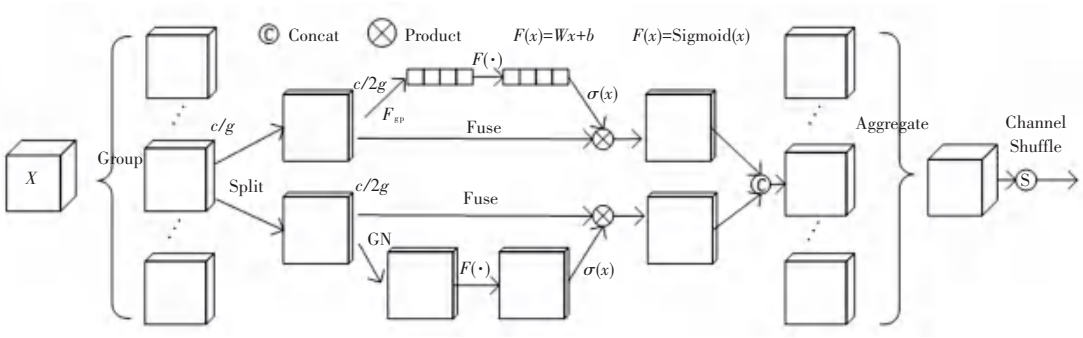

图 3　SA 注意力

Fig. 3　Shuffle Attention

随后，将注意力加权的结果在维度上进行重新融合，最终输出通过将融合后的特征图。

得益于 SA 注意力的作用，将其添加在 Backbone 的最后一层。该层是尺度内特征交互（AIFI）的输入 S5，这样能够更有效地获取深层次特征。改进残差模块结构如图 4 所示。

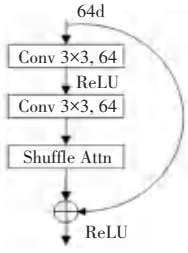

图 4　改进残差模块

Fig. 4　Improved residual module

2.2　添加 CGA 模块

级联分组注意力（Cascaded Group Attention，CGA）模块是 EfficientViT[24] 模型中的一个核心创新点，针对 Vision Transformer 模型的计算效率进行了优化。CGA 模块的核心思想是在自注意力机制中引入特征的多样性。CGA 模型设计如图 5 所示。与传统的自注意力机制不同，后者使用相同的特征给所有的 head 进行计算，CGA 为每个 head 提供不同的输入特征，然后级联这些 head 的输出特征。这种方法不仅减少了多头自注意力中的计算冗余，而且还通过增加网络深度来提高模型的处理容量。相较于标准的 MHSA，CGA 模块展现了更高的内存效率，使其在处理大规模视觉任务时更为高效。

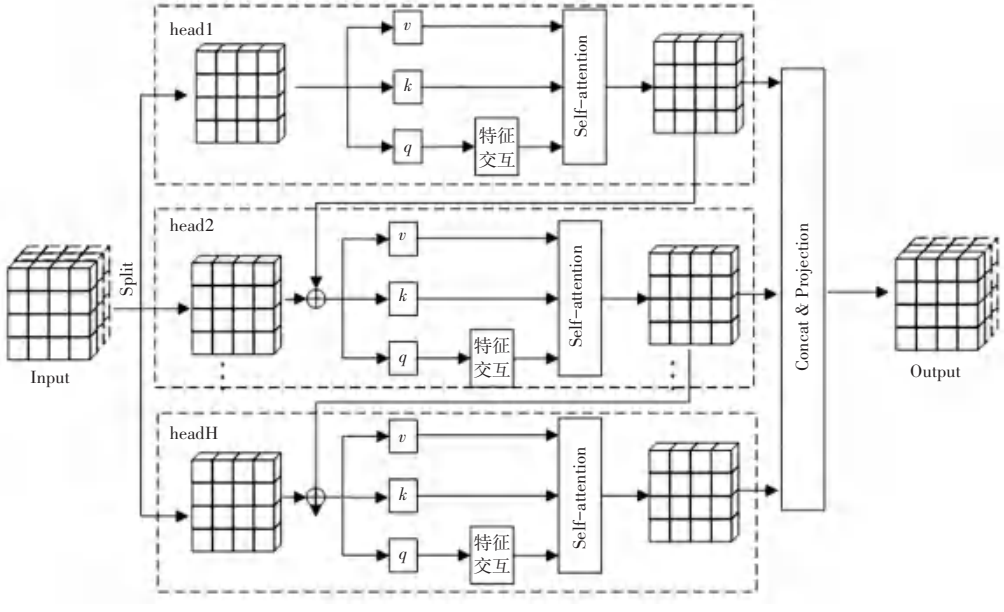

图 5　CGA 模块

Fig. 5　Cascaded Group Attention module

假设输入特征图为 X，维度为 $H \times W \times C$，其中 H 和 W 是空间维度，C 是通道数。将输入特征图平均分割成 N 个组，每组具有 C/N 个通道，将分割的通道进行自注意力计算。

对于每个头 $head_i$，特征图 X_i 经过线性变换生成 key $\boldsymbol{K}$、query $\boldsymbol{Q}$、value $\boldsymbol{V}$。研究推得公式如下：

$$\boldsymbol{K}_i = \boldsymbol{W}_i^K X_i \tag{5}$$

$$\boldsymbol{Q}_i = \boldsymbol{W}_i^Q X_i \tag{6}$$

$$\boldsymbol{V}_i = \boldsymbol{W}_i^V X_i \tag{7}$$

其中，$\boldsymbol{W}_i^K$，$\boldsymbol{W}_i^Q$，$\boldsymbol{W}_i^V$ 分别表示对应的权重矩阵。进一步又推得：

$$A_i = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{8}$$

其中，d_k 表示 key 的维度。研究中沿着 K 的列来应用 Softmax 函数。

在此基础上，应用自注意力得到特征图 $\boldsymbol{X}_i'$，即：

$$\boldsymbol{X}_i' = A_i \boldsymbol{V} \tag{9}$$

在后续计算中如果不是第一个 head，则将前一个 head 的输出与当前 head 的输入进行相加。具体公式如下：

$$\boldsymbol{X}_i = \boldsymbol{X}_i + \boldsymbol{X}_{i-1}' \tag{10}$$

最后，将所有 head 的输出特征图级联起来，通过一个线性变换投影回原始通道 C，由此可得到如下公式：

$$\boldsymbol{X}' = \mathrm{Concat}(\boldsymbol{X}_1', \boldsymbol{X}_2', \cdots, \boldsymbol{X}_N') \tag{11}$$

$$\mathrm{Output} = \boldsymbol{W}^P \boldsymbol{X}' \tag{12}$$

其中，$\boldsymbol{W}^P$ 表示输出投影的权重矩阵。

最后，输出特征图（Output）是经过级联分组注意力模块处理后的结果，其维度与输入特征图 X 相同，可以继续传递到网络的下一层或用于生成最终的任务特定输出。通过这种方式，CGA 模块允许不同 head 捕获不同的特征表示，最终通过级联获得一个综合的特征表示。

在级联结构中，每个注意力 head 的输出会作为下一个 head 的输入的一部分。这种递进式的处理方式意味着每个 head 不必从头开始处理完整的信息，而是在前一个 head 的结果基础上进一步提炼和增强特征，从而节约了计算资源。级联的方式可能还会减少需要学习的参数数量。虽然这不是 CGA 的直接目标，但参数量的减少往往会伴随着计算量的下降，从而对运算速度有帮助。在支持并行的计算设备上，因为每个 head 处理不同的特征组，这些操作可以并行执行，进一步提高计算效率。

将 CGA 模块引入到 RT-DETR 中，不仅降低了参数量和模型大小，并且最终的运算速度得到提升，对需要实时性的视觉 Transformer 模型十分友好。

3 实验设计与结果分析

3.1 数据集与实验环境

为了验证本文方法的有效性，本文在自建的车载驾驶环境数据集上进行了实验。该数据集专注于城市道路环境，通过在汽车驾驶平台中央放置手机进行视频拍摄来采集数据。文中所采集的图像具有 1 920×1 080 像素的分辨率，并以 25 帧/s 的帧率进行实验记录。通过每 7 s 抽取一帧的方式构建了车载驾驶环境数据集，并从中挑选了 1 600 张图像进行详细的数据标注。数据标注主要涵盖了图片中的城市道路环境参与者，如各种形式的车辆（vehicle）、骑行者（rider）和行人（pedestrian）。为了实验的需要，将这个数据集按照 7∶1∶2 的比例分为训练集和验证集和测试集。车载驾驶环境数据集样例如图 6 所示。

图 6 车载驾驶环境数据集样例

Fig. 6 Example of in-vehicle driving environment dataset

训练环境为：所使用环境操作系统为 Ubuntu 20.04 LTS，计算资源为 CPU i7-13700KF，1 张 NVIDIA RTX 4070ti 显卡，深度学习框架为 Pytorch 1.13.1。

训练时优化器参数设置为：优化器选用 AdamW，动量设置为 0.9，初始学习率为 0.000 1，关

闭混合精度训练(amp)，共训练 300 个 epochs。

3.2　评价指标

本文方法采用平均准确率(Average Precision，AP)、mAP(mean Average Precision)、FPS(Frames Per Second)和模型参数量对模型进行评价。

(1) 精确率(Precision)。是指正确预测为正确(TP)的占全部预测(TP+FP)的比例，计算公式为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

(2) 召回率(Recall)。是指正确预测为正确(TP)的占实际(TP+FN)的比例，计算公式为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

(3) AP。是某个类别的平均精确度。对于第 i 个类别，选取不同的 IoU 阈值，平均精确度的计算公式为：

$$\text{AP} = \int_0^1 \text{Precisiond}(\text{Recall}) \tag{15}$$

(4) mAP 。n 个类别的平均准确率计算公式如下：

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} \text{AP} \tag{16}$$

(5) FPS。是每秒检测图像帧数。该指标不仅与模型的参数量相关，还与实验过程中的硬件性能相关。

4　实验结果分析

为了验证本文改进的算法检测性能，选取参数量接近的常用目标检测算法 YOLO 系列和 Faster-RCNN 进行对比，使用数据集均为自建车载驾驶环境数据集。因为本文探讨的是实时性的目标检测算法，主要评判的是模型在实际检测过程中 FPS。

复杂道路场景实验结果对比如图 7 所示。模型在自建数据集上模型对比见表 1。在不同目标检测模型的比较中，基于 RT-DETR 的改进模型(RT-DETR-R18-SA-CGA)取得了 72.9%的平均精度(mAP)，显示了较为优秀的检测性能。这个结果比原始 RT-DETR-R18 模型的 71.8%有所提高。在每秒帧数(FPS)方面，改进模型达到了 131.0 FPS，超过了 FasterRCNN-R18、YOLOv5m 和 YOLOv8m 和原始 RT-DETR-R18 模型。消融实验见表 2。由表 2 可知，对 Baseline、Shuffle Attention(SA)、Cascaded Group Attention(CGA)的使用情况进行了对比。在使用 CGA 的实验组中，该模块能有效提高注意力模块的计算速度，SA 也能够提升检测性能。从实验结果可以看出，通过添加 SA 和 CGA 注意力机制，改进的 RT-DETR 模型在维持较高的检测精度的同时，还实现了在实时处理方面的性能提升。这说明了改进模型在处理动态和复杂的道路交通场景中的有效性和实用性。

(a) 原图

(b) RT-DETR-R18

(c) RT-DETR-R18-SA-CGA

图 7　复杂道路场景实验结果对比

Fig. 7　Comparison of experimental results of complex road scenes

表 1 模型在自建数据集上模型对比

Table 1 Model comparison on self-built datasets

Model	Input	mAP/%	模型大小/MB	参数量/M	FPS/(f·s^{-1})
FasterRCNN-r18	768×1344	77.3	221.0	28.6	42.3
YOLOv5m	768×1344	64.9	40.1	20.9	70.1
YOLOv8m	640×640	68.1	52.0	25.9	108.3
RT-DETR-r18	640×640	70.8	38.5	20.1	121.6
RT-DETR-r18-SA-CGA	640×640	**72.9**	**38.3**	**19.7**	**131.0**

表 2 消融实验

Table 2 Ablation experiment

实验组	Baseline	SA	CGA	0.5mAP/%	FPS
1	✓	×	×	71.8	121.6
2	✓	✓	×	72.3	121.3
3	✓	×	✓	70.8	132.5
4	✓	✓	✓	72.9	131.0

5 结束语

在本文中,通过在 RT-DETR 模型的骨干网络中添加 Shuffle Attention 机制,显著增强了对图像特征的捕获,提高了模型检测性能,经过尺度内特征交互模块的处理,这些信息进一步提升了模型的检测精度。此外,通过 Cascaded Group Attention 模块对 Transformer 的计算过程进行优化,有效提高了模型的检测速率。在专门设计的道路交通场景数据集上的性能评估显示,相较于其他模型,改进的 RT-DETR 模型在准确率和速度上均表现出色,证明了其在道路交通场景中实时应用的潜力。然而,由于 RT-DETR-R18 是最小的基线模型,本研究在低成本硬件或低算力设备上的应用可能会受限,且在不支持并行处理的设备上性能可能有所下降。未来的工作将探索优化模型以提高其在资源受限环境下的可用性。

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Long Beach, USA: NIPS Foundation, 2017: 5998-6008.

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

[3] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

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

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

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

[7] LI Chuyi, LI Lulu, JIANG Hongliang, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. arXiv preprint arXiv, 2209.02976, 2022.

[8] 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. Piscataway, NJ: IEEE, 2023: 7464-7475.

[9] TIAN Zhi, SHEN Chunhua, CHEN Hao, et al. FCOS: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 9627-9636.

[10] GE Zheng, LIU Songtao, WANG Feng, et al. YOLOx: Exceeding YOLO series in 2021[J]. arXiv preprint arXiv, 2107.08430, 2021.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv, 2010.11929, 2020.

[12] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012-10022.

[13] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer,

2020：213-229.
[14] ZHU Xizhou，SU Weijie，LU Lewei，et al. Deformable detr：Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv，2010.04159，2020.
[15] LIU Shilong，LI Feng，ZHANG Hao，et al. Dab-detr：Dynamic anchor boxes are better queries for DETR[J]. arXiv preprint arXiv，2201.12329，2022.
[16] ZHANG Hao，LI Feng，LIU Shilong，et al. Dino：Detr with improved denoising anchor boxes for end-to-end object detection[J]. arXiv preprint arXiv，2203.03605，2022.
[17] LV Wenyu，ZHAO Yi′an，XU Shangliang，et al. Detrs beat yolos on real-time object detection[J]. arXiv preprint arXiv，2304.08069，2023.
[18] 王鹏飞，黄汉明，王梦琪. 改进 YOLOv5 的复杂道路目标检测算法[J]. 计算机工程与应用，2022，58(17)：81-92.
[19] 盛博莹，侯进，李嘉新，等. 面向复杂交通场景的道路目标检测方法[J]. 计算机工程与应用，2023，59(15)：87-96.
[20] 冉险生，苏山杰，陈俊豪，等. 自适应特征融合的复杂道路场景目标检测算法[J]. 计算机工程与应用，2023，59(24)：216-226.
[21] 李轩，李静，王海燕. 密集交通场景的目标检测算法研究[J]. 计算机技术与发展，2020，30(7)：46-50.
[22] 李永上，马荣贵，张美月. 改进 YOLOv5s+DeepSORT 的监控视频车流量统计[J]. 计算机工程与应用，2022，58(5)：271-279.
[23] ZHANG Qinglong，YANG Yubin. SA-Net：Shuffle attention for deep convolutional neural networks[C]//Proceedings of 2021 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). Piscataway，NJ：IEEE，2021：2235-2239.
[24] LIU Xinyu，PENG Houwen，ZHENG Ningxin，et al. EfficientViT：Memory efficient vision transformer with cascaded group attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2023：14420-14430.
[25] HE Kaiming，ZHANG Xiangyu，REN Shaoqing，et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2016：770-778.