·智能车辆运动规划与控制技术专题·

基于Transformer改进的YOLOv5+DeepSORT的车辆跟踪算法*

何水龙[1,2]　张靖佳[1]　张林俊[1]　莫德赟[2]

（1.桂林电子科技大学，桂林 541004；2.桂林航天工业学院，桂林 541001）

【摘要】针对传统目标检测跟踪算法检测精度低、全局感知能力差、对遮挡和小目标物体的识别能力差等问题，提出了一种基于轻量化Transformer改进的YOLOv5和DeepSORT算法的车辆跟踪方法。首先，利用EfficientFormerV2模型改进YOLOv5算法模型，增强车辆的目标检测能力；然后，利用移位窗口（Swin）模型的优点改进DeepSORT多目标跟踪算法中的重识别（Re-Identification）模块，提高车辆的跟踪能力和精度；最后，通过数据集KITTI和VeRi开展对比试验和消融实验。结果表明，在复杂工况下，该方法的性能在车辆遮挡和小目标识别方面显著提高，平均准确度达到96.7%，目标跟踪准确度提高了9.547%，编号（ID）切换总次数减少了26.4%。

主题词：YOLOv5　车辆检测　DeepSORT　Transformer

中图分类号：TP391.41；U463.6　　**文献标志码：**A　　**DOI:** 10.19620/j.cnki.1000-3703.20231097

Vehicle Tracking Algorithm Based on Transformer's Improved YOLOv5+DeepSORT

He Shuilong[1,2], Zhang Jingjia[1], Zhang Linjun[1], Mo Deyun[2]

（1. Guilin University of Electronic Technology, Guilin 541004; 2. Guilin University of Aerospace Technology, Guilin 541004）

【Abstract】 In order to solve the shortcomings of traditional object detection and tracking algorithms, such as low detection accuracy, poor global perception ability, poor recognition ability of occlusion and small target objects, this paper proposed a vehicle tracking method based on YOLOv5 and DeepSORT algorithm improved by lightweight Transformer. Firstly, the EfficientFormerV2 model was used to improve the YOLOv5 algorithm model to enhance the target detection ability of the vehicle, and then the advantages of the Swin model were used to improve the Re-Identification module in the DeepSORT multi-target tracking algorithm to enhance the tracking ability and accuracy of the vehicle. Finally, the dataset KITTI and VeRi were used to carry out comparative experiments and ablation experiments. The results show that under complex conditions, the performance of the proposed method is significantly improved in vehicle occlusion and small target recognition, with an average accuracy of 96.7%, an increase of 9.547% in target tracking, and a reduction of 26.4% in the total number of ID switching.

Key words: YOLOv5, Vehicle detection, DeepSORT, Transformer

【引用格式】 何水龙，张靖佳，张林俊，等. 基于Transformer改进的YOLOv5+DeepSORT的车辆跟踪算法[J]. 汽车技术，2024(7)：9-16.

HE S L, ZHANG J J, ZHANG L J, et al. Vehicle Tracking Algorithm Based on Transformer's Improved YOLOv5+DeepSORT[J]. Automobile Technology, 2024(7): 9-16.

1　前言

目标识别和跟踪技术是提高高级辅助驾驶系统安全性能的核心手段之一，其通过实时识别并跟踪车辆、行人和道路标志等目标，帮助车辆感知周围交通状况，减少交通事故。

近年来，深度学习在目标检测领域不断发展。2017年，He等[1]提出了掩膜循环卷积神经网络（Mask Recycle

*基金项目：广西科技重大专项（AA22068001，AA23062031）；广西重点研发项目（AB21196029）；柳州市科技计划项目（2022AAA0102）。

通信作者：莫德赟（1983—），男，副教授，主要研究方向为汽车智能驾驶，23440217@qq.com。

Convolutional Neural Network，Mask R-CNN）算法，有效解决了原图与特征图的特征位置不匹配的问题。2018年，Redmon等[2]在改进基础网络的同时，结合金字塔结构，提出了YOLOv3[3]算法，获取了更多小目标的有效信息。2019年，Zhao等[4]针对目标尺度变化的问题，提出了M2Det算法。2020年后，基于YOLOv3改进的YOLOv4[5]和YOLOv5[6]模型在保持运行效率优势的基础上提高了检测与识别的准确率。然而，这些方法在某些方面仍然存在一定的局限性，如：Mask R-CNN在实现上比快速循环卷积神经网络（Faster Recycle Convolutional Neural Network，Faster R-CNN）[7]复杂，需要更多的计算资源，且使用了类似于Faster R-CNN的两阶段目标检测方法，检测速度相对较慢；YOLO系列模型在处理小目标和遮挡目标时仍存在挑战；M2Det算法需要处理多个尺度的特征金字塔，故其在实时性上并不理想。

随着深度学习技术的发展，多目标跟踪算法也不断改进。Yu等[8]提出了一个两阶段算法，先使用Faster R-CNN进行目标检测，再利用匈牙利算法对由GoogleNet[9]提取的特征进行关联，从而实现目标跟踪。Xie等[10]利用基于YOLOv3的检测器捕捉目标，并使用DeepSORT（Deep learning based Simple Online and Realtime Tracking）算法实现轨迹关联。然而，两阶段算法需要两个密集计算网络，存在跟踪效率低的问题。因此，诸多研究者转向基于重识别（Re-IDentification，Re-ID）技术的多目标跟踪算法研究，提高多目标跟踪效率。Wang等[11]率先提出了一种联合模型，通过改进YOLOv3检测模型，一次性解决目标检测和Re-ID特征提取，在行人数据集上实现了较高水平的跟踪效率。Zhang等[12]提出了FairMOT算法，使用深层特征融合网络进行特征提取，从而提高了跟踪性能。但上述算法所使用的骨干网络都是由检测器网络改造而来，在学习Re-ID特征上存在缺陷。

为进一步提升目标跟踪算法精度、效率和跟踪能力，本文提出一种基于轻量化Transformer改进的YOLOv5和DeepSORT的车辆跟踪方法，弥补YOLO系列对于小目标和遮挡物的检测能力不足以及DeepSORT中Re-ID模块泛化能力弱的缺点。

2 目标检测算法

2.1 YOLOv5算法模型

YOLOv5是一种基于深度残差和路径聚合网络的目标检测算法，其骨干网络基于CSPDarknet53[13]，结合特征金字塔网络（Feature Pyramid Networks，FPN）[14]和空间金字塔池化（Spatial Pyramid Pooling，SPP）[15]技术，提升了小目标检测精度。在COCO数据集[16]上，YOLOv5的平均精度均值（mean Average Precision，mAP）表现优异[17-18]，超越了当时的最先进水平。

2.2 YOLOv5算法改进

YOLOv5采用CSPDarknet53对输入数据进行划分，通过拆分路由（Split Route）模块分为两个部分，然后用跨阶段部分（Cross Stage Partial，CSP）模块连接，再通过一个大卷积层将特征融合，从而得到骨干网络输出的特征图。这种操作能够很好地处理图像的局部特征。然而，由于YOLOv5采用无锚点（Anchor-Free）方式，在单个目标的检测方面存在缺陷。如在小目标物体检测和物体被遮挡的情况下，存在检测漏报和误报的情况。针对这种情况，本文提出一种改进YOLOv5目标检测模型，如图1所示。该模型在保证网络正常检测较大目标的同时，提高对小目标特征信息的感知能力和全局感知能力，以提高遮挡物体的识别率和泛化能力，满足实时性和提高检测精度的要求，采用最新的轻量化Transformer模型EfficientFormerV2[19]对YOLOv5的骨干网络进行改进。EfficientFormerV2使用全局自注意力机制，在处理道路交通领域的车辆目标检测任务时，特别是在存在大量背景干扰的情况下，能够有效地分割不同区域对应的目标对象，达到更好的检测效果。采用快速空间金字塔池化（Spatial Pyramid Pooling-Fast，SPPF）模块连接EfficientFormerV2模块，在不同尺度的特征图中划分多个子区域，并利用最大池化对每个子区域进行处理。最终将所有尺度的池化结果拼接成一个固定长度的特征向量，解决不同尺度特征图的融合问题，在处理车辆遮挡和全局感知方面可获得更好的效果。

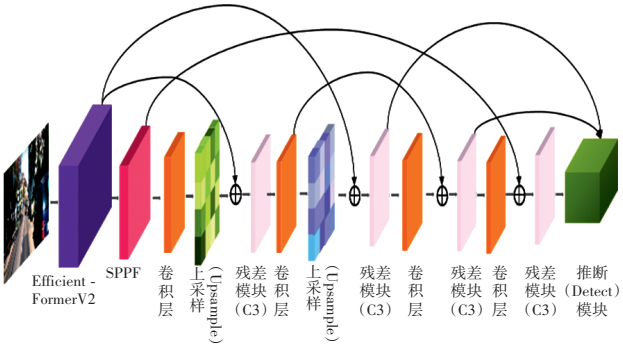


图1 改进后YOLOv5的网络模型框架

2.3 EfficientFormerV2网络模型

EfficientFormerV2是Detransformer模型的改进版，基于Transformer的自注意力机制，能有效处理对象关系与局部图像信息，其网络结构如图2所示。本文选用其轻量化版本EfficientFormerV2-S2，参数量仅10.3×10^6个，适用于边缘计算处理器部署。

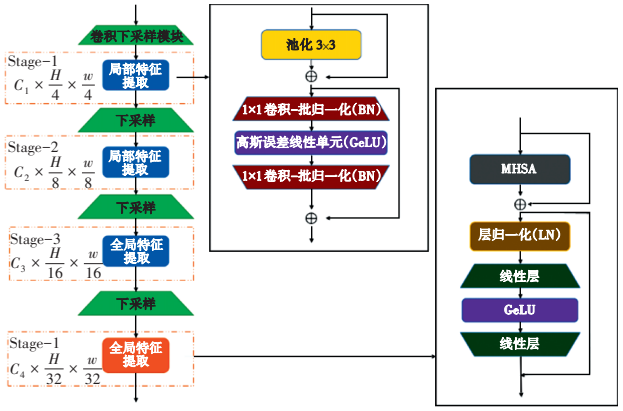


图2　EfficientFormerV2的网络结构

EfficientFormerV2采用了四阶段分层设计，可以获得输入图像分辨率在{1/4,1/8,1/16,1/32}处的特征图。为更高效地嵌入输入图像，EfficientFormerV2使用了小内核卷积，而不是非重叠补丁（Patch）的方式，从而提高了计算性能和模型泛化能力。该设计使得EfficientFormerV2在图像分类和目标检测等任务中都获得了极佳的性能表现。计算过程为：

$$X_{i\,|_{i=1}\,j|_{j=1}}^{B,C_{j|_{j=1}},\frac{H}{4},\frac{W}{4}} = stem\left(\chi_0^{B,3,H,W}\right) \tag{1}$$

式中：$X_{i,j}$表示第i层第j阶段的特征图，$j\in\{1,2,3,4\}$，B为批大小，C_j为第j阶段通道大小（表示网络宽度），H、W分别为特征图的高度和宽度，χ_0为输入图像，$stem$为卷积下采样操作。

第一阶段和第二阶段的设计旨在以高分辨率捕获局部信息，采用了相同的前馈神经网络（Feedforward Neural Network，FFN）来处理每层特征图，如图3所示。这种设计使得EfficientFormerV2能够在局部区域获取更多的细节信息，有助于实现更准确的目标检测和图像分类：

$$X_{i+1,j}^{B,C_j,\frac{H}{2^{i+1}},\frac{W}{2^{i+1}}} = S_{i,j}\cdot \mathrm{FFN}^{C_j,E_{i,j}}\left(X_{i,j}\right) + X_{i,j} \tag{2}$$

式中：$S_{i,j}$为一种可学习的层间尺度；FFN含有两种属性，即阶段宽度C_j和每块扩展比$E_{i,j}$。

图3　前馈神经网络

需要注意的是，每个FFN都采用了残差连接（Residual Connection）。在模型的最后两个阶段，本地FFN和全局多头自注意力（Multi-Head Self-Attention，MHSA）块均被使用。

本文将4个FFN模块封装在一个时序（Sequential）容器中，可方便地对它们进行堆叠和复用，避免手动重复编码。此外，在第2层、第4层、第6层的时序容器与批标准化（Batch Normalization）结合使用。其中，时序容器对输入的序列进行局部特征提取和非线性变换，而批归一化则可以对每个时序容器模块的输出进行标准化处理，减少数据内部协方差的影响，从而加速模型收敛并降低过拟合风险。EfficientFormerV2模块的输出特征向量被传递给SPPF模块和下游的其他卷积层。SPPF模块通过网络池化操作生成固定长度的特征向量，用于下游任务。

3　目标跟踪算法

3.1　DeepSORT算法

简单在线实时跟踪（Simple Online and Realtime Tracking，SORT）[20]利用卡尔曼滤波器预测目标运动，通过交并比（Intersection Over Union，IOU）评估预测边界框与检测边界框的相似度，并应用匈牙利算法关联数据，实现实时跟踪。DeepSORT在SORT基础上引入深度学习网络提取目标特征，采用级联匹配技术解决目标重叠或遮挡时的编号（ID）切换问题。该算法结合运动与外观特征计算代价矩阵，匹配检测结果，将未匹配的目标视为新目标，分配新ID。级联匹配技术根据目标丢失次数和轨迹活跃程度对目标进行优先排序，有效减少了ID切换次数。

3.2　DeepSORT算法改进

目标特征提取的主要目的是获得目标的唯一标识特征，以便对其在不同位置或姿态下进行重新识别，从而实现目标跟踪。在DeepSORT算法中，特征提取的主要算法是基于卷积神经网络（Convolutional Neural Network，CNN）的ResNet-50[21]，用以对目标图像区域进行卷积特征提取。对于每个检测目标，先裁剪其位置，再经CNN提取卷积特征，通过全连接层降维得到特征向量。该向量反映目标视觉与外观信息，鲁棒性强，不受位置和姿态变化的影响。ResNet-50在ImageNet上进行了大规模预训练，故提取的特征向量更准确且区分力更强。

不过，ResNet-50也存在一定不足：首先，ResNet-50具有非常深的网络结构，导致训练和推理速度较慢，尤

其是在高分辨率图像上；其次，ResNet-50的感受野较大，当目标物体较小时，容易忽略一些关键信息，导致检测失败；最后，由于ResNet-50相对于一些轻量级神经网络而言体积更大，需要更多的存储和计算资源。为解决这些问题，本文对DeepSORT中的重识别模块进行了改进，将ResNet-50主干网络换成基于Transformer架构的移位窗口（Shifted windows，Swin）[22]，如图4所示。Swin凭借分布式训练、跨群组部署及计算与存储分离等优势，可实现快速训练和推理，并展现出较强的可扩展性。其分级特征提取与多重注意力机制使得小目标检测敏感度超越了ResNet-50。计算注意力机制相似度时，在每个头（Head）中加入相对位置偏置$\boldsymbol{B} \in \mathbf{R}^{M^2 \times M^2}$：

$$\text{Atention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{SoftMax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d} + \boldsymbol{B}\right)\boldsymbol{V} \quad (3)$$

式中：$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} \in \mathbf{R}^{M^2,d}$分别为查询（Query）矩阵、键（Key）矩阵和价值（Value）矩阵，d为查询矩阵、键矩阵的维度，M^2为局部窗口内的补丁数量。

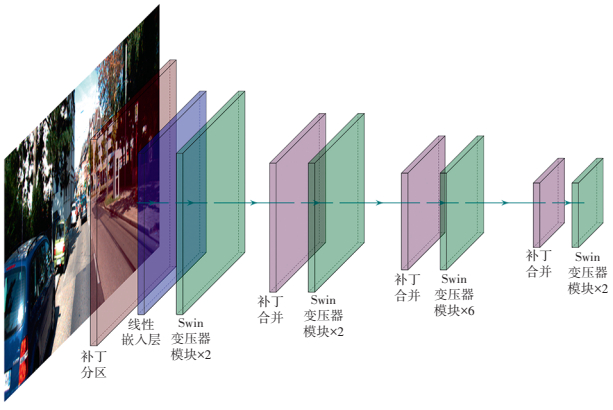


图4　基于Swin Transformer改进后网络模型架构

此外，该算法还提出了横向和纵向的多重特征信息响应，这种分层设计的思路不仅方便根据任务调整网络深度，而且可以有效避免梯度消失等问题。

4　试验结果

本文采用仅有27 MB的轻量化YOLOv5s模型，兼顾精度、速度与成本，提升算法运行性能。

4.1　试验配置

本文试验采用开源的PyTorch深度学习框架。CPU使用第12代Intel Core i7-12700H，主频为4.70 GHz；采用Ubuntu20.04 LTS操作系统，其中包含Python 3.8和CUDA 12.0；图形处理器使用GeForce GTX 3060，显存容量为6 GB。

为适配KITTI数据集，本文对YOLOv5进行了重新训练，优化了训练参数与批大小（Batch Size），如表1所示，并利用文献[19]开源的权重加速收敛。

表1　试验参数配置

参数名称	YOLOv5	EfficientFormerV2
权重文件	Yolov5s.pth	eformer_s2_450.pth
代数	100	100
批大小	6	6
初始学习率	0.01	0.01
动量	0.937	0.937
预设衰减系数	0.000 5	0.000 5

4.2　数据集

采用KITTI数据集[23]对模型进行测试和评估，KITTI数据集作为自动驾驶与计算机视觉评估的核心基准，包含多序列多视角图像数据。针对其与YOLOv5模型的不兼容性，本研究进行了预处理：数据被细分为六类目标，格式转为xml，并适配为YOLOv5训练标签，从而推进其在该模型中的有效应用。

VeRi车辆重识别数据集[24]是用于研究车辆重识别的公共数据集之一。该数据集涵盖20种摄像机视角下的视频及576辆车共计37 778张图像，展现多视角、多样图像质量（含模糊、噪声），及车辆局部细节（如车牌、车灯），适用于车辆重识别训练与算法性能评估。

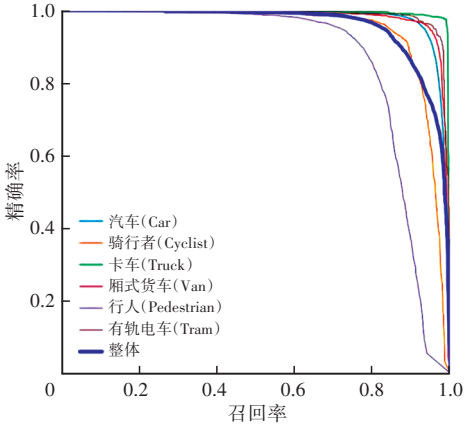
4.3　改进YOLOv5试验结果和分析

4.3.1　定量分析

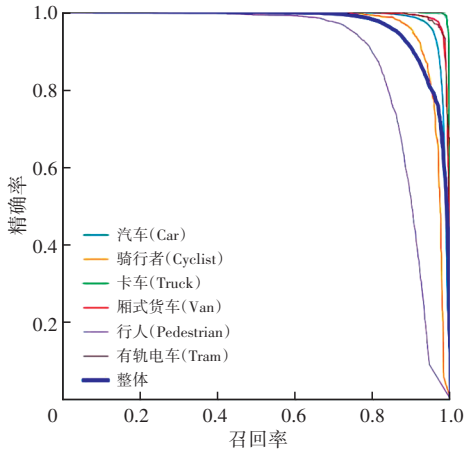
图5所示为改进YOLOv5算法的对比试验结果。可以看出，改进算法在IOU阈值为0.5时的mAP明显提高，从95.6%提升至96.7%，说明了本文的方法能够有效提高对车辆目标的检测能力。

4.3.2　定性分析

算法定性试验结果如图6所示，改进前的算法明显未能识别右下角的红色汽车，而改进后的算法成功地识别了该车辆。试验结果表明，改进后的YOLOv5具备更强的全局感知能力，对于车辆目标跟踪具有更好的泛化性能。



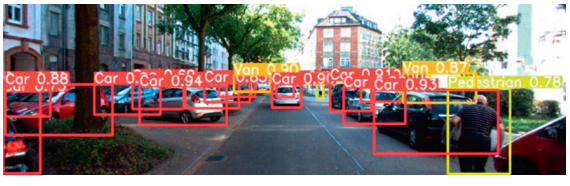
（a）改进前



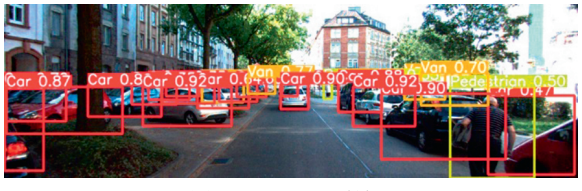
(b)改进后

图5　改进前、后的YOLOv5算法试验结果对比

(a)原始图片



(b)改进前的识别效果



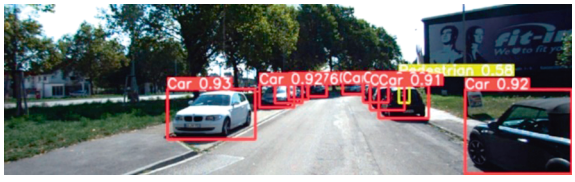
(c)改进后的识别效果

图6　全局目标识别效果对比

算法对遮挡物体的识别效果如图7所示。图7中，道路右侧前方的黑色轿车挡住了行人。改进前的算法无法识别被遮挡行人，而改进后的算法则能够正确识别。因此，改进后的YOLOv5在物体遮挡识别方面表现出色。

(a)原始图片

(b)改进前的识别效果



(c)改进后的识别效果

图7　遮挡物体识别效果对比

为了验证改进后算法的小目标检测效果，进行了相关试验，结果如图8所示，由于YOLOv5对于识别小目标准确度比较低，并未识别到小目标行人，而改进算法成功识别到目标。试验对比结果表明，改进算法对于小目标的检测能力显著提高。

(a)原始图像

(b)改进前的识别效果

(c)改进后的识别效果

图8　小目标识别效果对比

4.3.3　改进前、后性能对比

KITTI数据集每个目标的标注行都包含了截断(Truncated)字段，表示相应物体在图像中是否被边界框截断，其取值通常在0~1范围内，表示目标相对于实际规模的截断程度。这个信息对于理解物体在图像中的完整性和全局性非常重要，尤其是在自动驾驶场景下。

试验计算了整个数据集中不同截断程度下的目标数量，为4 631个，并分成了多个段位，如图9所示。通过计算改进前、后算法中数据集内不同截断程度的目标识别成功数量，进而形成了改进前、后的效果对比。可以看出，截断程度越大，识别成功率越低，但改进算法成功识别的数量明显比原算法更多，充分说明改进算法在全局感知能力上有较好的提升效果。

遮挡(Occluded)属性通常表示物体被其他物体遮挡的程度，在KITTI标注中，该属性的值为整数。取值包括：

0表示物体没有被遮挡，即物体在图像中是完全可见的；1表示物体被部分遮挡；2表示物体被大部分遮挡，但仍然可见；3表示物体被完全遮挡，即物体在图像中不可见。

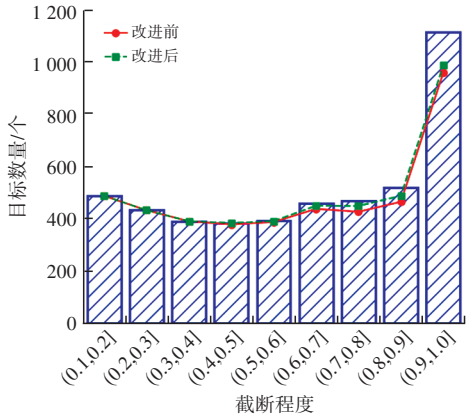


图9　全局感知能力效果对比

根据数据集的标注属性统计了不同遮挡程度的目标总数，如图10所示。从试验统计结果可以看出，改进算法的识别成功数量明显比原算法的数量多，特别是在大部分遮挡的情况下，改进算法比原算法识别成功率高12.8%。

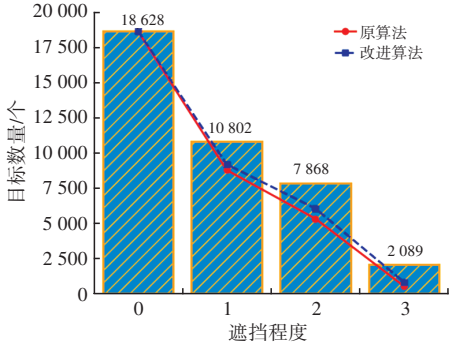


图10　遮挡物体识别效果对比

根据COCO数据集对于小目标的定义，本文采用相同策略，将32×32以下像素点的目标定义为小目标，符合小目标要求的总数量为6 756个。

通过试验结果可以看出，原算法的小目标识别率为84.9%，改进算法的识别率为92.82%，如图11所示，可以看出，改进算法在识别小目标上有明显优势。

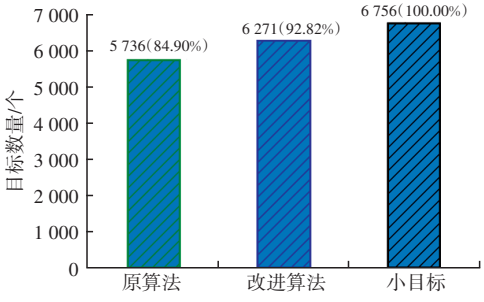


图11　小目标识别效果对比

根据试验结果可知，相较于原算法，改进后的YOLOv5算法改善了全局感知能力，提高了遮挡物的检测和小目标的识别效果，同时提升了目标检测的准确率。

4.4　改进DeepSORT试验结果和分析

针对重识别模块的模型对比试验，本文使用了基于开源代码DeepSORT的重识别模型。由于DeepSORT模型中默认使用ResNet-50作为网络模型，将其替换为Swin Transformer，并保持初始化参数相同，试验结果如表2所示。可见，改进模型的平均精度提升了8.13%，Rank-1精度（Rank-1 Accuracy）提升了3.35%。说明Transformer模型增强了传统CNN模型的多尺度特征融合能力，能够更好地提取多尺度特征，从而提高识别的准确率。

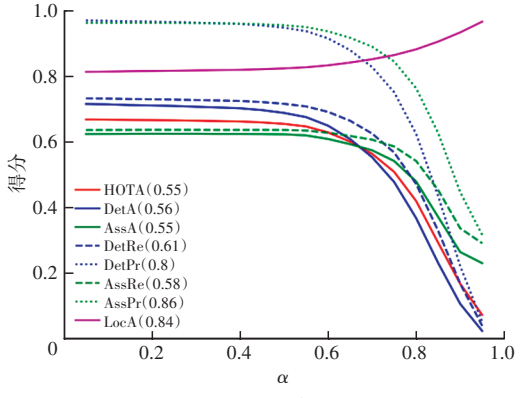
表2　DeepSORT改进试验对比　%

项目	ResNet-50	Swin Transformer
mAP	71.56	79.69
Rank-1精度	88.83	92.18

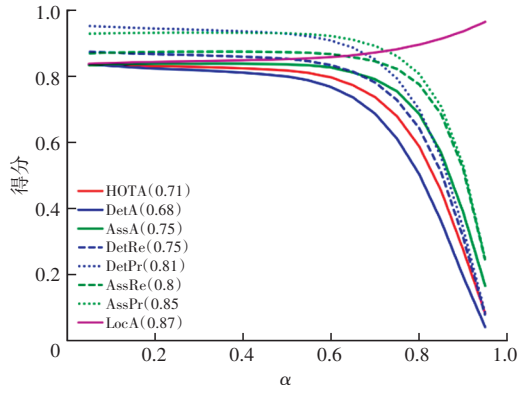
上述结果说明了算法模型改进的有效性。本文将改进后的算法应用于YOLOv5s+DeepSORT，并与原算法进行对比，高阶跟踪精度（Higher Order Tracking Accuracy，HOTA）、检测精确度（Detection Accuracy，DetA）、关联精确度（Association Accuracy，AssA）、检测精度（Detection Precision，DetPr）、关联召回（Association Recall，AssRe）、关联精度（Association Precision，AssPr）、定位精度（Localization Accuracy，LocA）结果如图12所示。其中，α为权衡因子，用于平衡定位（LocA）、关联（AssA、AssRe、AssPr）和检测（DetA、DetPr）之间的关系，α越大，表示更重视关联和检测的性能，α越小，表示更侧重于定位的精度。由图12可知，改进算法在HOTA指标上明显提高，从55%提升至71%，表明将主干网络从CNN改变为Transformer对于模型性能具有积极影响。

4.5　消融实验

为了进一步验证所提出算法的检测性能，探究各改进方法的有效性，在YOLOv5s+DeepSORT的基础上设计了3组消融实验，每组实验使用相同的超参数以及训练技巧，实验结果如表3所示。



(a)改进前



(b)改进后

图12　HOTA指标对比

表3　不同方法评估结果

算法	MOTA/%	ID变换次数/次
YOLOv5+DeepSORT	69.137	460
改进YOLOv5+DeepSORT	77.105	424
YOLOv5+改进DeepSORT	67.723	405
改进YOLOv5+改进DeepSORT	78.684	339

消融实验结果表明，改进后的YOLOv5在识别准确度方面显著提升，能够将多目标跟踪准确度（Multiple Object Tracking Accuracy，MOTA）提升7.968百分点并降低ID变换总次数。虽然改进后的DeepSORT在精度上有所损失，MOTA降低了1.414百分点，但ID变换总次数下降了12%，表明改进的重识别能够有效提取目标特征，并具有对姿态、遮挡和光照等方面的鲁棒性。最终改进版比原始版本在目标跟踪准确度上提高了9.547%，ID切换总次数减少了26.4%。因此，在DeepSORT中，计算特征之间相似度的准确度得到了提高，从而导致ID转换频率的降低。

4.6　跟踪试验验证

本文基于KITTI数据集，验证了改进后目标跟踪算法的有效性，该算法在处理小目标和遮挡物体时性能更优秀，同时具备更强的全局感知能力。试验结果如图13所示，改进后的算法表现更加出色。

(a)改进前小目标识别效果

(b)改进后小目标识别效果

(c)改进前遮挡识别效果

(d)改进后遮挡识别效果

(e)改进前全局识别效果

(f)改进后全局识别效果

图13　目标跟踪算法改进前、后识别效果对比

5　结束语

本文提出了一种基于改进YOLOv5和DeepSORT的车辆检测及跟踪算法。使用轻量化网络EfficientFormerV2替换了原YOLOv5模型的主干网络CSPDarknet53，在减少模型参数的同时提取到了更多潜在的特征信息，提高了特征的代表性。在跟踪阶段，DeepSORT算法中的重识别网络结构也得到了优化，通过增加正则化和利用Swin Transformer网络模型重新设计网络主干技术，进一步提高了外观信息提取能力和跟踪能力。试验结果表明，该方法在公共数据集上取得了更优的检测和跟踪效果，目标跟踪准确度提高了9.547%，ID切换总次数减少了26.4%。

本文所构建的目标跟踪方法除在交通安全和智慧交通等领域具有研究价值外，也可为其他目标检测和跟踪任务提供新的思路和方法。但该方法未能实现端到端的目标跟踪，在未来的研究中，可以考虑在轻量化Transformer基础上实现端到端的跟踪，以进一步提高跟踪算法的性能。

参　考　文　献

[1] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on

Computer Vision. Venice, Italy: IEEE, 2017: 2961-2969.

[2] TAN M X, LE Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]// International Conference on Machine Learning. Long Beach, California: PMLR, 2019: 6105-6114.

[3] SHEN L Z, TAO H F, NI Y Z, et al. Improved YOLOv3 Model with Feature Map Cropping for Multi-Scale Road Object Detection[J]. Measurement Science and Technology, 2023, 34(4).

[4] ZHAO Q J, SHENG T, WANG Y T, et al. M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, Hawaii, USA: AAAI, 2019: 9259-9266.

[5] YU J M, ZHANG W. Face Mask Wearing Detection Algorithm Based on Improved YOLO-v4[J]. Sensors, 2021, 21(9): 3263.

[6] WU W T, LIU H, LI L L, et al. Application of Local Fully Convolutional Neural Network Combined with YOLO v5 Algorithm in Small Target Detection of Remote Sensing Image[J]. PLoS One, 2021, 16(10).

[7] BHARATI P, PRAMANIK A. Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey[C]// Computational Intelligence in Pattern Recognition. Singapore: Springer, 2020: 657-668.

[8] YU F W, LI W B, LI Q Q, et al. POI: Multiple Object Tracking with High Performance Detection and Appearance Feature[C]// Computer Vision-ECCV 2016 Workshops. Cham, Switzerland: Springer, 2016: 36-42.

[9] YU Z G, DONG Y Y, CHENG J H, et al. Research on Face Recognition Classification Based on Improved GoogleNet[J]. Security and Communication Networks, 2022, 2022.

[10] 谢金龙，胡勇．基于深度学习的车辆检测与跟踪系统[J]. 工业控制计算机，2020，33(7)：99-101.
XIE J L, HU Y. Vehicle Detection and Tracking System Based on Deep Learning[J]. Industrial Control Computer, 2020, 33(7): 99-101.

[11] WANG Z D, ZHENG L, LIU Y X, et al. Towards Real-Time Multi-Object Tracking[C]// European Conference on Computer Vision. Cham, Switzerland: Springer, 2020: 107-122.

[12] CHE J, HE Y T, WU J M. Pedestrian Multiple-Object Tracking Based on FairMOT and Circle Loss[J]. Scientific Reports, 2023, 13(1): 4525.

[13] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, WA, USA: IEEE, 2020: 390-391.

[14] HE K M, ZHANG X Y, REN S Q, 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.

[15] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[16] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common Objects in Context[C]// 13th European Conference on Computer Vision. Zurich, Switzerland: Springer International Publishing, 2014: 740-755.

[17] REDMON J, FARHADI A. YOLOv3: An Incremental Improvement[EB/OL]. (2018-04-08)[2024-01-18]. https://arxiv.org/abs/1804.02767.

[18] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[EB/OL]. (2020-04-23)[2024-01-18]. https://arxiv.org/abs/2004.10934.

[19] LI Y Y, HU J, WEN Y, et al. Rethinking Vision Transformers for MobileNet Size and Speed[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023.

[20] BEWLEY A, GE Z Y, OTT L, et al. Simple Online and Realtime Tracking[C]// 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA: IEEE, 2016: 3464-3468.

[21] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.

[22] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 10012-10022.

[23] GEIGER A, LENZ P, STILLER C, et al. Vision Meets Robotics: The KITTI Dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[24] LIU X C, LIU W, MA H D, et al. Large-Scale Vehicle Re-Identification in Urban Surveillance Videos[C]// 2016 IEEE International Conference on Multimedia and Expo (ICME). Seattle, WA, USA: IEEE, 2016: 1-6.

（责任编辑　斛　畔）

修改稿收到日期为2024年1月18日。