激光与光电子学进展 先进成像

# 融合YOLO-V4与改进SiameseRPN的多目标跟踪算法

朱志玲[1]，周志峰[1*]，赵勇[2]，王永泉[3]，王立端[3]

[1]上海工程技术大学机械与汽车工程学院，上海 201620；

[2]山西中电科新能源技术有限公司，山西 太原 030024；

[3]上海司南卫星导航技术股份有限公司，上海 201801

**摘要** 针对现有多目标跟踪算法精度不高的问题，提出了一种融合YOLO-V4与改进SiameseRPN的多目标跟踪算法。首先通过YOLO-V4网络自动获取跟踪目标，制作模板后输入SiameseRPN跟踪网络；然后在模板分支中采用背景自适应策略初始化模板，并且融合残差连接构建Siamese网络；最后通过匈牙利算法对YOLO-V4的检测结果和改进SiameseRPN的跟踪结果进行数据关联，实现多目标跟踪。实验结果表明，与其他算法相比，所提算法具有较好的跟踪性能，在目标尺度变化、外观变化、部分遮挡等情况下能够实现稳定跟踪。

**关键词** 机器视觉；多目标跟踪；SiameseRPN算法；背景自适应；数据关联

**中图分类号** TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.2215010

# Multiobject Tracking Algorithm Combining YOLO-V4 and Improved SiameseRPN

Zhu Zhiling[1], Zhou Zhifeng[1*], Zhao Yong[2], Wang Yongquan[3], Wang Liduan[3]

[1]*School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai* 201620, *China*;

[2]*Shanxi New Energy Technology Co., Ltd., Taiyuan* 030024, *Shanxi*, *China*;

[3]*Shanghai Compass Satellite Navigation Technology Co., Ltd., Shanghai* 201801, *China*

**Abstract** This study proposes a multiobject tracking algorithm combining YOLO-V4 and improved SiameseRPN to overcome the low accuracy of existing multiobject tracking algorithms. First, the tracking objects are automatically obtained using the YOLO-V4 network. After creating the template, enter it into the SiameseRPN tracking network. Then, the adaptive background strategy is adopted in the template branch to initialize the template, and the Siamese network is constructed by integrating residual connections. Finally, the results of YOLO-V4 and improved SiameseRPN are used to perform data association through the Hungarian algorithm to achieve multiobject tracking. The experimental results show that the proposed algorithm has better tracking performance than other algorithms. Furthermore, the proposed algorithm can achieve stable tracking under object scale, appearance change, and partial occlusion conditions.

**Key words** machine vision; multiobject tracking; SiameseRPN algorithm; background adaptation; data association

## 1 引 言

多目标跟踪[1]是计算机视觉领域的重要分支，广泛应用于行人监测、自动驾驶、安防等领域，具有重要的实用价值。但在实际应用场景中，存在目标尺度变化、外观形变、遮挡、目标进出视野等干扰，这些因素导致跟踪精度大大受限。

传统的基于检测的背景差分法[2]、光流法[3]由于计算量较大，实时性较差；而帧差法[4]、阈值分割法[5]虽然跟踪速度较快，但易受环境光、图像噪声、形变的影响，造成精度低；经典的生成式模型粒子滤波[6]、Meanshift[7]、Camshift[8]等算法实时性较强，但在目标快速运动情况下易产生误差累积，使目标发生漂移，严重影响目标的跟踪精度。近年来，随着深度学习的迅速

**收稿日期**：2021-08-19；**修回日期**：2021-10-07；**录用日期**：2021-10-19

**基金项目**：上海市科学技术委员会科研基金(17511106700)

**通信作者**：*zhousjtu@126.com

发展，深度神经网络极大地促进了目标跟踪的发展，众多结合目标检测网络的跟踪算法开始涌现。SiameseRPN[9]将Siamese网络与目标检测中的区域候选网络（RPN）融合，采用RPN更精准预测跟踪目标的位置，实现了精度与速度的提升，但该算法易受相似特征背景的干扰，且无法进行多目标跟踪。王殿伟等[10]提出了一种基于改进SiameseRPN的目标跟踪算法，该算法通过MobileNetV3提取特征，一定程度上提高了算法的适应性，但仍然存在定位不够准确的问题。陈法领等[11]结合VGGNet-19与相关滤波算法实现跟踪，提高了算法在目标形变、光照变化等复杂情况下的跟踪稳定性，但在遮挡条件下的跟踪精度有所下降。为了提高对背景的判别能力，金立生等[12]采用Gaussian YOLO-V3优化DeepSort跟踪算法，但该算法并没有提升运动目标的跟踪精度。上述算法虽然取得了较好的效果，但在复杂环境下无法实现高精度、实时的多目标跟踪。

针对上述问题，本文提出了一种融合YOLO-V4与改进SiameseRPN的多目标跟踪算法。该算法将目标检测模型YOLO-V4与跟踪模型SiameseRPN相结合，利用YOLO-V4自动获取跟踪目标，然后采用自适应背景初始化方法提高模板特征的判别性，通过残差连接优化孪生网络来增强特征表达，最后对检测与跟踪结果进行数据关联匹配，实现多目标跟踪。所提算法在保证跟踪速度的前提下，有效提升了跟踪精度，实现了稳定的多目标跟踪。

## 2　所提算法

基于深度学习的目标检测算法YOLO-V4[13]首先采用自对抗训练和Mosaic方式进行数据增广，通过新的特征网络CSPDarknet53增强模型的学习能力，在检测过程中，将提取的特征映射输入空间金字塔池化（SPP）模块，增大感受野，然后利用PANet结构融合特征，通过YOLO层进行分类和回归，得到检测结果。该算法实现了速度与精度的平衡，在复杂场景下检测能力较强，背景信息利用率高，具有优秀的定位精度，可适用于不同尺度目标的检测任务。

SiameseRPN由Siamese和RPN组成，Siamese部分的模板与搜索分支均采用AlexNet前五层提取图像特征，将所提特征输入RPN的分类与回归分支进行预测。在后续帧的跟踪中，通过计算模板特征与搜索区域特征之间的相似度响应图得到目标位置。该算法虽然具有良好的跟踪性能，但没有充分利用背景信息，导致在复杂场景下存在判别力不足的问题。在进行连续跟踪时，需要人工选定跟踪目标，无法自动获取，且目标变化时会导致跟踪失败。

YOLO-V4网络可弥补SiameseRPN模型的不足，因此可通过集成学习构建检测跟踪一体化模型，提升跟踪精度。具体模型框架如图1所示：首先通过YOLO-V4自动获取视频序列首帧中的跟踪目标，将该帧的结果作为模板图像$Z$输入SiameseRPN；然后在模板分支中采用自适应背景初始化策略，充分利用模板帧的先验信息来提高目标的判别能力；同时结合残差连接设计了13层的轻量网络提取特征，增强特征的表达能力；然后将不同分支的特征图输入RPN中进行相似度计算获得跟踪结果；最后，对YOLO-V4的检测结果和改进SiameseRPN的跟踪结果进行数据关联匹配，实现多目标跟踪。

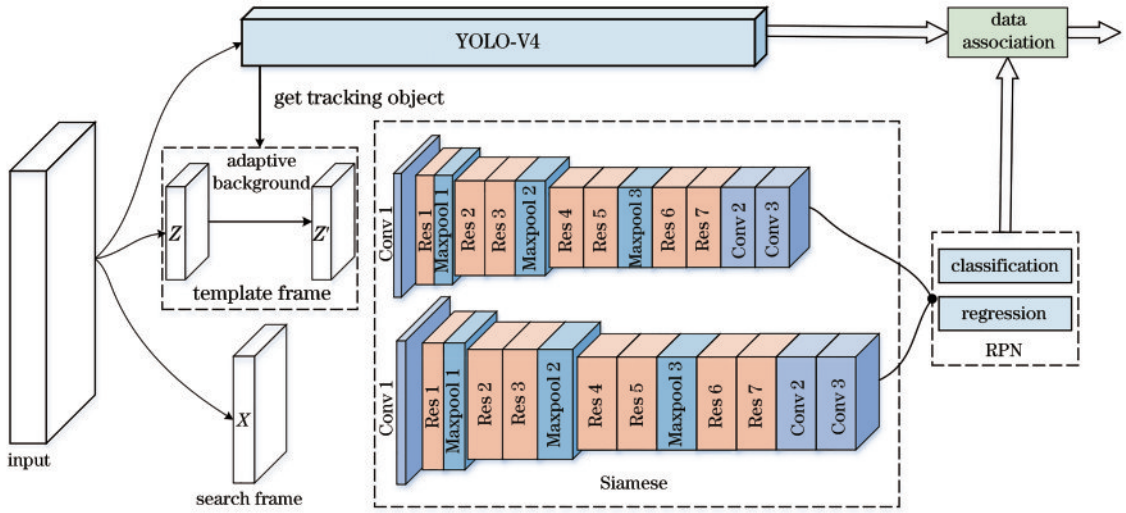


图1　所提算法模型框架

Fig. 1　Overall framework of proposed algorithm

### 2.1　融合跟踪模型

#### 2.1.1　自动获取目标

SiameseRPN在跟踪之前需要在视频首帧中人为自行选定跟踪目标，在乘客、行人跟踪等任务中，这种在视频端人工标识的方式人力成本过高，不仅浪费时间、效率低，而且在长时间的跟踪中存在由于目标发生

变化而导致的跟踪失败现象。针对此问题，所提融合模型利用目标检测的高精度、高可靠性的检测优势，通过 YOLO-V4 网络准确获取视频首帧的待跟踪目标。在后续帧的跟踪中，若出现新的目标或旧目标消失，可根据检测与跟踪的匹配结果更新待跟踪目标。

其主要原理如下：首先对视频序列进行预处理，分割出视频首帧图像，并将分割出的第 1 帧图像输入 YOLO-V4 模型中；通过此检测网络获取图像中目标的位置，得到 $N$ 个检测框；然后由这 $N$ 个目标制作模板图像，完成跟踪器初始化；在后续跟踪过程中由检测与跟踪的匹配结果判断是否更新模板。

#### 2.1.2 数据关联

SiameseRPN 网络广泛应用于单目标跟踪，利用其跟踪能力与速度良好的优点，所提融合模型通过构建 YOLO-V4 与 SiameseRPN 的并联结构来实现多目标跟踪。当 YOLO-V4 获取首帧目标位置后，将各目标分别作为模板，由 SiameseRPN 实现各目标的跟踪，通过 YOLO-V4 和 SiameseRPN 对视频后续帧同时进行检测跟踪。为了实现多目标跟踪，本实验组采用匈牙利算法[14]进行数据关联匹配，将 YOLO-V4 检测得到的检测框与 SiameseRPN 获取的跟踪框关联匹配。对于匹配成功的目标，便认为检测得到的目标位置为当前帧的跟踪结果。

根据匈牙利算法的思想，在关联匹配过程中，首先计算 YOLO-V4 网络得到的目标检测框与 SiameseRPN 获取的目标跟踪框之间的交并比（IoU）[15]值，然后将得到的 IoU 值组成状态关联矩阵，根据设定的 IoU 阈值评价检测框与跟踪框的匹配程度，成功匹配的检测框为最优结果并赋予目标 ID。

本实验组采用 IoU 来度量跟踪框与检测框之间的运动匹配程度，IoU 阈值作为超参数对数据关联的匹配效果有着一定的影响，不同的 IoU 阈值对检测框和跟踪框的匹配结果都会产生一定的数值变动，这可能导致算法出现误配和错检。因此，选择合适的阈值至关重要。第 $i$ 个跟踪框与第 $j$ 个检测框之间的 IoU 的表达式为

$$R_{\mathrm{IoU}_{i,j}}=\frac{S_i\cap S_j}{S_i\cup S_j},\tag{1}$$

式中：$S_i$ 为第 $i$ 个跟踪框的面积；$S_j$ 为第 $j$ 个检测框的面积。IoU 阈值一般在 0.5～0.7 之间，通过大量对比实验，本实验组最终将数据关联过程中的阈值设置为 0.55。

### 2.2 融合残差连接

在深层卷积神经网络中，不同特征层所包含的特征信息侧重点不同。深层特征含有丰富的高语义信息，有利于实现较高置信度分类；浅层特征则具有更丰富的位置、轮廓等强定位信息，有利于定位任务。而目标跟踪任务既需要深层语义特征判别正负样本，同时又需要丰富的强定位特征实现目标的精准定位。根据问题的复杂性，通常通过增加网络深度或宽度来提升模型性能，然而增加宽度的代价往往远高于深度。网络加深使得模型具有更强大的表达能力和逐层的特征学习能力。更深的模型意味着更好的非线性表达能力，可以学习更加复杂的变换，从而拟合更加复杂的特征输入。另外，网络更深，每一层需要学习的变换更加简单，而网络如果只有一层，就意味着要学习的变换非常复杂，这很难做到。因而，SiameseRPN 的孪生网络部分仅采用 AlexNet 的前五层卷积进行特征提取，获得的特征信息丰富度不足，在一定程度上影响了跟踪效果，难以应对复杂的跟踪环境。而 ResNet50、VGG16 等结构更深的网络会极大影响算法的跟踪速度。基于此，为提高跟踪精度，所提算法在 AlexNet 的基础上添加残差结构适当加深网络，通过提高网络特征提取能力并融合深浅层信息，使得网络在尽量维持实时性的同时增强模型的表征性能。改进后，具体的网络结构如表 1 所示。

由表 1 可知，网络中有 3 个 Conv 层、7 个 Res 层和 3 个最大池化层，卷积均采用 3×3 或 1×1 的卷积核。网络中多次采用了 1×1 的卷积，主要用于通道数降维，减小网络参数量，能够缓解网络加深引起的速度下降问题。同时，1×1 的卷积增加了模型的非线性表征能力，提高了网络的鲁棒性。残差结构在增加网络深度的同时加强了前后层特征信息的流通，使得特征具有更强的表达性。

### 2.3 自适应背景初始化

在 SiameseRPN 算法中，可能由于跟踪目标不匹配问题导致模板发生漂移，这种漂移现象在后续跟踪过程中通常会逐渐累积，从而影响模型精度。SiameseRPN 是利用孪生网络进行相似性度量实现目标跟踪的，因此模板图像的特征信息尤为重要，它是后续帧中目标相似性度量的基准。针对上述问题，本实验组采用自适应背景初始化的策略来增强目标特征的差异性表达，弱化背景的干扰。通常来说，前景与背景的差异性越大，模型就越容易辨别出目标。所提背景自适应初始化算法能够增大前景与背景的对比度，从而提高模板分支对跟踪目标特征的判别能力。

模板图像分为目标和背景：如果两者像素均值的差值小于某阈值且目标的像素均值大于 127，则用 0 填充背景像素值；如果两者像素均值的差值小于某阈值且目标的像素均值小于 127，则用 255 填充背景像素值；其他情况采用原模板图像。背景初始化方式可描述为

$$B_{\mathrm{color}}=\begin{cases}0, & \left|M_{\mathrm{t}}-M_{\mathrm{b}}\right|\leqslant T_{\mathrm{c}}\ \text{and}\ M_{\mathrm{t}}\geqslant 127\\ 255, & \left|M_{\mathrm{t}}-M_{\mathrm{b}}\right|\leqslant T_{\mathrm{c}}\ \text{and}\ M_{\mathrm{t}}<127\\ \text{remain the same}, & \text{otherwise}\end{cases},\tag{2}$$

式中：$M_{\mathrm{t}}$、$M_{\mathrm{b}}$ 分别为目标与背景的像素均值；$T_{\mathrm{c}}$ 表示阈值。通过此策略对原始模板图像进行背景变换，然

表 1 融合残差连接的 Siamese 网络结构

Table 1 Siamese network structure incorporating residual connections

| Convolution layer | Convolution kernel | Output and input channel | Stride | Template image | Search image | Channel |
|---|---|---|---|---|---|---|
| | | | | 127×127 | 255×255 | 3 |
| Conv1 | 3×3 | 64×3 | 1 | 125×125 | 253×253 | 64 |
| Res1 | 1×1 | 64×64 | 1 | 123×123 | 251×251 | 32 |
| | 3×3 | | | | | 64 |
| Maxpool1 | 2×2 | | 2 | 61×61 | 125×125 | 64 |
| Res2 | 1×1 | 128×64 | 1 | 59×59 | 123×123 | 64 |
| | 3×3 | | | | | 128 |
| Res3 | 1×1 | 128×128 | 1 | 57×57 | 121×121 | 64 |
| | 3×3 | | | | | 128 |
| Maxpool2 | 2×2 | | 2 | 28×28 | 60×60 | 128 |
| Res4 | 1×1 | 256×128 | 1 | 26×26 | 58×58 | 128 |
| | 3×3 | | | | | 256 |
| Res5 | 1×1 | 256×256 | 1 | 24×24 | 56×56 | 128 |
| | 3×3 | | | | | 256 |
| Maxpool3 | 2×2 | | 2 | 12×12 | 28×28 | 256 |
| Res6 | 1×1 | 256×256 | 1 | 10×10 | 26×26 | 128 |
| | 3×3 | | | | | 256 |
| Res7 | 1×1 | 512×256 | 1 | 8×8 | 24×24 | 256 |
| | 3×3 | | | | | 512 |
| Conv2 | 1×1 | 256×512 | 1 | 8×8 | 24×24 | 256 |
| Conv3 | 3×3 | 256×256 | 1 | 6×6 | 22×22 | 256 |

后将调整后的模板图像输入 Siamese 网络获取模板特征以增强判别能力。

### 2.4 算法流程

所提算法在改进 SiameseRPN 跟踪模型的基础上与 YOLO-V4 网络融合，实现准确高效的多目标跟踪。算法流程如图 2 所示，具体的步骤如下：

1）输入视频序列，利用 YOLO-V4 网络获取视频首帧跟踪目标的位置；

2）使用第 1 帧的检测结果作为模板图像，通过自适应背景初始化算法完成跟踪器初始化；

3）利用 YOLO-V4 网络检测当前帧目标，得到检测框集合 $N$，采用改进 SiameseRPN 跟踪各目标，获得跟踪框集合 $M$；

4）计算 YOLO-V4 网络和改进 SiameseRPN 得到

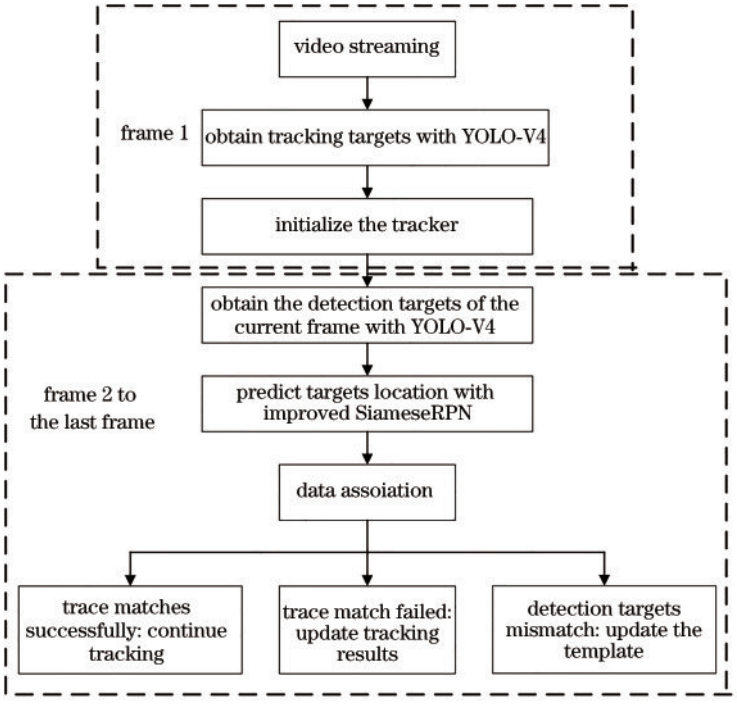


图 2 多目标跟踪算法流程

Fig. 2 Multitarget tracking algorithm flow

的目标位置集合$N$与$M$的IoU值，获得IoU关联矩阵；

5）当$N$与$M$数量相等且IoU大于设定阈值时，匹配成功，模型继续进行后续跟踪；

6）当跟踪框与检测框的IoU值不大于设定阈值时，不匹配，将检测结果作为此帧的目标位置；

7）当跟踪框数量少于检测框时，表示出现新目标，此时根据检测结果建立新模板图像；

8）重复上述步骤至跟踪结束。

## 3　实验结果与分析

所提算法基于python语言和pytorch框架实现，实验操作系统为Ubuntu18.04，内存为32 GB，CPU为Intel Xeon(R)W-2135，GPU为GeForce RTX2080Ti，显存为10 GB。

本实验组通过采集地铁闸机场景下的乘客通行图像自建了乘客检测与跟踪数据集。为验证所提算法的有效性和泛化能力，在自建数据集上训练YOLO-V4网络，采用改进SiameseRPN分别在VOT2016和VOT2018数据集上进行对比实验，在OTB100数据集上对各改进策略进行消融实验，并最终在自建数据集上验证所提算法的多目标跟踪性能。

### 3.1　检测器训练实验

多目标跟踪实验前需要训练YOLO-V4网络，使用自建的乘客数据集进行训练。自建的乘客检测数据集包含18000张图像，训练集、验证集和测试集比例为4∶1∶1。初始学习率为0.001，batch size为8，共训练300个epoch。训练过程损失值和平均精度(AP)变化曲线如图3所示。

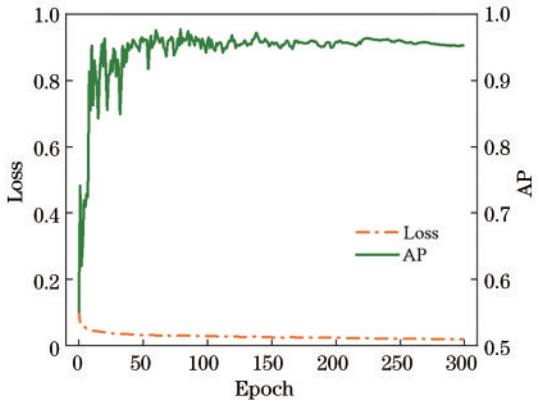


图3　损失与平均精度曲线

Fig. 3　Loss and AP curves

训练结果表明，YOLO-V4网络在自建的乘客检测数据集上的精确率和召回率分别为88.5%、96.4%，平均精度为95.6%，检测速度为62 frame·s$^{-1}$，达到了良好的检测效果。图4为YOLO-V4网络的部分检测结果，从图4可以看出，所提算法在尺度不同、目标进出视野的情况下仍能保证良好的检测性能。

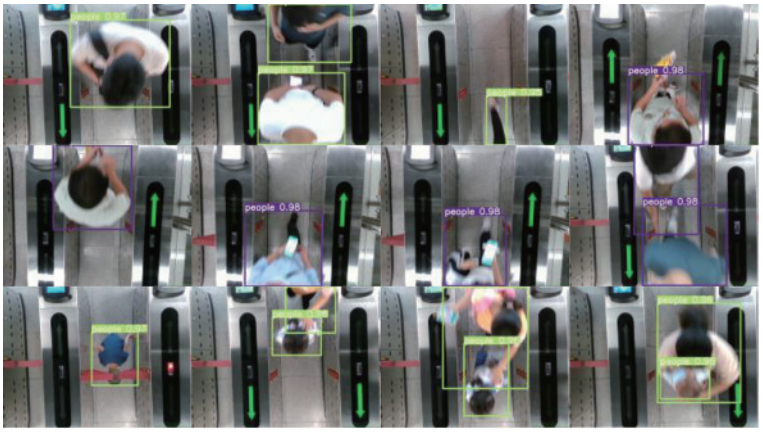

图4　乘客检测结果

Fig. 4　Passenger detection results

### 3.2　改进SiameseRPN跟踪实验

VOT2016和VOT2018是近年跟踪领域广泛使用的数据集，分别含有60组视频序列，主要采用准确性(Acc)、鲁棒性(Rob)和平均重叠期望(EAO)进行算法评估与对比。跟踪模型的准确性与平均重叠期望越高、鲁棒性值越低，表明其综合性能越好。实验中采用恺明正态分布初始化网络，在GOT-10k上进行预训练。初始学习率设为0.01，采用指数衰减将学习率衰减至$10^{-5}$，衰减系数设为0.0005。批大小为8，通过SGD优化，共训练50个epoch。余弦窗口影响因子设为0.27。

#### 3.2.1　VOT2016实验结果

为验证所提改进SiameseRPN算法的有效性，基于VOT2016数据集与其他主流跟踪算法进行了对比实验，包括SiameseFC[16]、Staple[17]、SRDCF[18]、DeepSRDCF[19]、TADT[20]和SiameseRPN，其实验结果如表2所示。由表2可知，所提改进SiameseRPN算法的准确性、平均重叠期望指标最高，鲁棒性指标最低，相较于原算法准确性提高了1.91个百分点，鲁棒性提高了10.6个百分点，平均重叠期望提高了6.53个百分点，表明所提算法具有优异的跟踪性能。图5(a)和

表 2　多种算法在 VOT2016 上的对比

Table 2　Comparison results of different algorithms on VOT2016 dataset

| Algorithm | Acc | Rob | EAO |
|---|---|---|---|
| SiameseFC | 0.5342 | 0.4613 | 0.2356 |
| Staple | 0.5425 | 0.3784 | 0.2946 |
| SRDCF | 0.5356 | 0.4193 | 0.2459 |
| DeepSRDCF | 0.5271 | 0.3264 | 0.2758 |
| TADT | 0.5539 | 0.3327 | 0.2991 |
| SiameseRPN | 0.5617 | 0.2621 | 0.3442 |
| Proposed algorithm | 0.5808 | 0.1561 | 0.4095 |

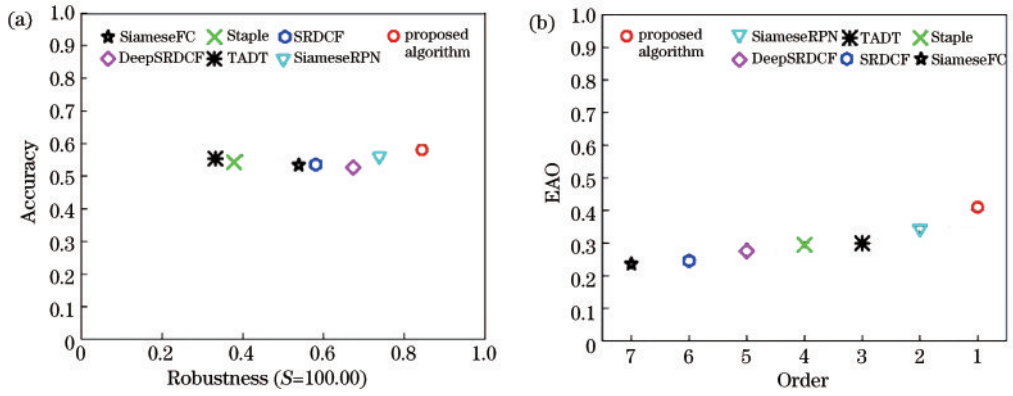


图 5　VOT2016 上的性能指标排名。(a)精度-鲁棒性;(b)平均重叠期望

Fig. 5　Performance indicator ranking on VOT2016. (a) Accuracy-robustness; (b) average overlap expectation

图 5(b)分别为在 VOT2016 测试集上的精度-鲁棒性排名与平均重叠期望排名,所提算法排名第 1,说明所提算法总体跟踪性能最好,在复杂环境下具有良好的稳定性。

### 3.2.2　VOT2018 实验结果

所提算法基于 VOT2018 数据集与 SiameseFC、Staple、DSiamese[21]、SiameseDW[22]、CFNet[23]和 SiameseRPN 进行了对比实验,其实验结果如表 3 所示。由表 3 可知,所提算法的准确性略低于 SiameseDW,排名第 2,平均重叠期望值最高,鲁棒性值最低,与 SiameseRPN 相比,所提算法的准确性提高了 4.02 个百分点,鲁棒性提高了 12.1 个百分点,平均重叠期望提高了 6.5 个百分点,说明所提算法跟踪性能优异。图 6(a)和图 6(b)分别为在 VOT2018 测试集上的精度-鲁棒性排名与平均重叠期望排名,所提算法综合排名较好,即所提算法具有良好的跟踪性能和泛化能力,在不同环境下能够实现较好的稳定性。

表 3　多种算法在 VOT2018 上的对比

Table 3　Comparison results of different algorithms on VOT2018 dataset

| Algorithm | Acc | Rob | EAO |
|---|---|---|---|
| SiameseFC | 0.5108 | 0.4836 | 0.2343 |
| Staple | 0.5246 | 0.6887 | 0.1694 |
| DSiamese | 0.5117 | 0.6458 | 0.1966 |
| SiameseDW | 0.5411 | 0.4032 | 0.2704 |
| CFNet | 0.5028 | 0.5853 | 0.1882 |
| SiameseRPN | 0.4945 | 0.4627 | 0.2441 |
| Proposed algorithm | 0.5347 | 0.3417 | 0.3091 |

### 3.2.3　消融实验

为验证所提融合残差连接和自适应背景初始化优化策略的有效性,在 OTB100 数据集上通过比较不同优化策略对算法产生的效果来评估其对跟踪性能的影响。使用跟踪成功率(AUC)和跟踪精度(Prec)作为评价指标,实验结果如表 4 所示,其中 SiameseRPN+RC 为采用融合残差连接设计的模型,SiameseRPN+AB 为利用自适应背景初始化方法改进的模型,SiameseRPN+RC+AB 为采用两种策略改进的模型。

由表 4 可知,融合残差连接和自适应背景初始化策略在 AUC 和 Prec 指标上均有一定程度的提高,表明所提两种优化策略均能有效提高 SiameseRPN 算法的性能。

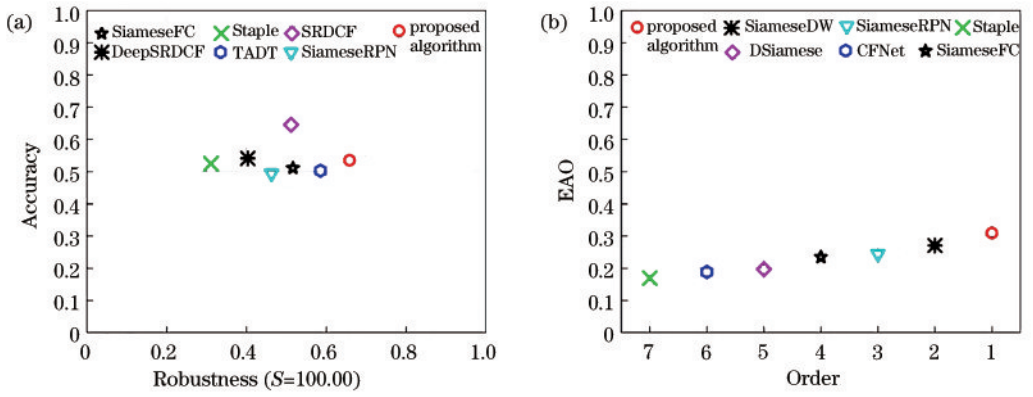


图 6　VOT2018 上的性能指标排名。(a)精度-鲁棒性;(b)平均重叠期望

Fig. 6　Performance indicator ranking on VOT2018. (a) Accuracy-robustness; (b) average overlap expectation

表 4　在 OTB100 上的消融实验结果

Table 4　Experimental results of ablation on OTB100 dataset

| Algorithm | AUC | Prec |
|---|---|---|
| SiameseRPN | 0.596 | 0.785 |
| SiameseRPN+RC | 0.638 | 0.822 |
| SiameseRPN+AB | 0.609 | 0.804 |
| SiameseRPN+RC+AB | 0.654 | 0.846 |

#### 3.2.4　定量分析

OTB100 数据集包含不同视频属性:尺度变化(SV)、目标遮挡(OCC)、形变(DEF)、超出视野(OV)、背景干扰(BC)等,本实验组在该数据集中将所提算法与其他跟踪算法在各难点属性视频序列中进行对比,比较各算法在不同场景下的表现。表 5 为各算法在这些复杂跟踪因素下的平均 Prec。

表 5　各算法在 OTB100 上对不同视频属性的 Prec 对比

Table 5　Comparison of Prec of different video attributes on OTB100 for each algorithm

| Attribute | Number of videos | SiameseFC | Staple | SRDCF | ACFN | SiameseRPN | Proposed algorithm |
|---|---|---|---|---|---|---|---|
| SV | 63 | 0.765 | 0.721 | 0.739 | 0.758 | 0.806 | 0.841 |
| OCC | 48 | 0.738 | 0.729 | 0.730 | 0.752 | 0.781 | 0.837 |
| DEF | 43 | 0.779 | 0.742 | 0.725 | 0.764 | 0.793 | 0.843 |
| OV | 14 | 0.695 | 0.670 | 0.593 | 0.687 | 0.736 | 0.780 |
| BC | 31 | 0.703 | 0.747 | 0.767 | 0.766 | 0.803 | 0.824 |

从表 5 可以看出,所提算法在 OTB 数据集的 5 个复杂场景下性能最优,在应对上述场景时,较 SiameseRPN 在 Prec 上均有一定程度的提升,表明所提算法具有较好的鲁棒性,跟踪性能更稳定。

### 3.3　多乘客跟踪实验

自建的乘客跟踪数据集包含 10 组视频序列,每组时长 10 min 左右,数据集中包含了目标形变、尺度变化、部分遮挡、超出视野等常见问题的样本。实验中采用多目标跟踪精确度(MOTP)、多目标跟踪准确度(MOTA)和每秒帧率(FPS)作为评价指标。MOTP 表示跟踪预测框与真实框之间的平均匹配程度,MOTA 衡量了模型保持连续跟踪能力和跟踪一致性的性能,FPS 反映算法跟踪的实时性。为验证所提融合 YOLO-V4 与改进 SiameseRPN 的多目标跟踪算法的跟踪性能,在自建数据集上进行了实验,并与 SORT[24]、DeepSort[25]、MHT[26]、POI[27]和 SiameseCNN[28]算法进行了对比,实验结果如表 6 所示。

由表 6 可知,所提算法应用于地铁通行乘客跟踪时精确度为 89.1%,准确度为 93.7%,相比排名第 2 的 POI 算法,MOTP 提升了 5.6 个百分点,MOTA 提升了 5.4 个百分点,并且具有实时性。实验结果表明,本实验组利用检测网络优秀的定位精度与改进 SiameseRPN 网络较强的跟踪能力的优点,实现了性能更好的多目标跟踪算法。

表 6　各算法性能对比

Table 6　Performance comparison of each algorithm

| Algorithm | MOTP | MOTA | FPS |
|---|---|---|---|
| SORT | 0.814 | 0.815 | 112 |
| DeepSort | 0.817 | 0.869 | 48 |
| MHT | 0.822 | 0.871 | 1.5 |
| POI | 0.835 | 0.883 | 15 |
| SiameseCNN | 0.764 | 0.865 | 107 |
| Proposed algorithm | 0.891 | 0.937 | 39 |

图 7 为采用所提算法与 POI 算法在乘客跟踪数据集上的实验对比结果:图 7(a)中当目标开始进入或出视野时,POI 算法不能很好地预测目标,容易出现目标

丢失，而所提算法能够持续跟踪目标；图 7(b)中 POI 算法由于背景颜色干扰产生跟踪漂移，而所提算法能够始终锁定目标位置；图 7(c)中所提算法在目标被部分遮挡时仍有较好的表现，而 POI 跟踪失败；图 7(d)和图 7(e)中目标存在形变和尺度变化时，POI 算法跟踪时会发生漂移，丢失位置精度，而所提算法有更准确的位置信息表达。由上述分析可知，所提算法在目标进出视野、形变、部分遮挡等情况下能够更准确稳定地跟踪目标，自适应跟踪能力更好，具有较好的鲁棒性。

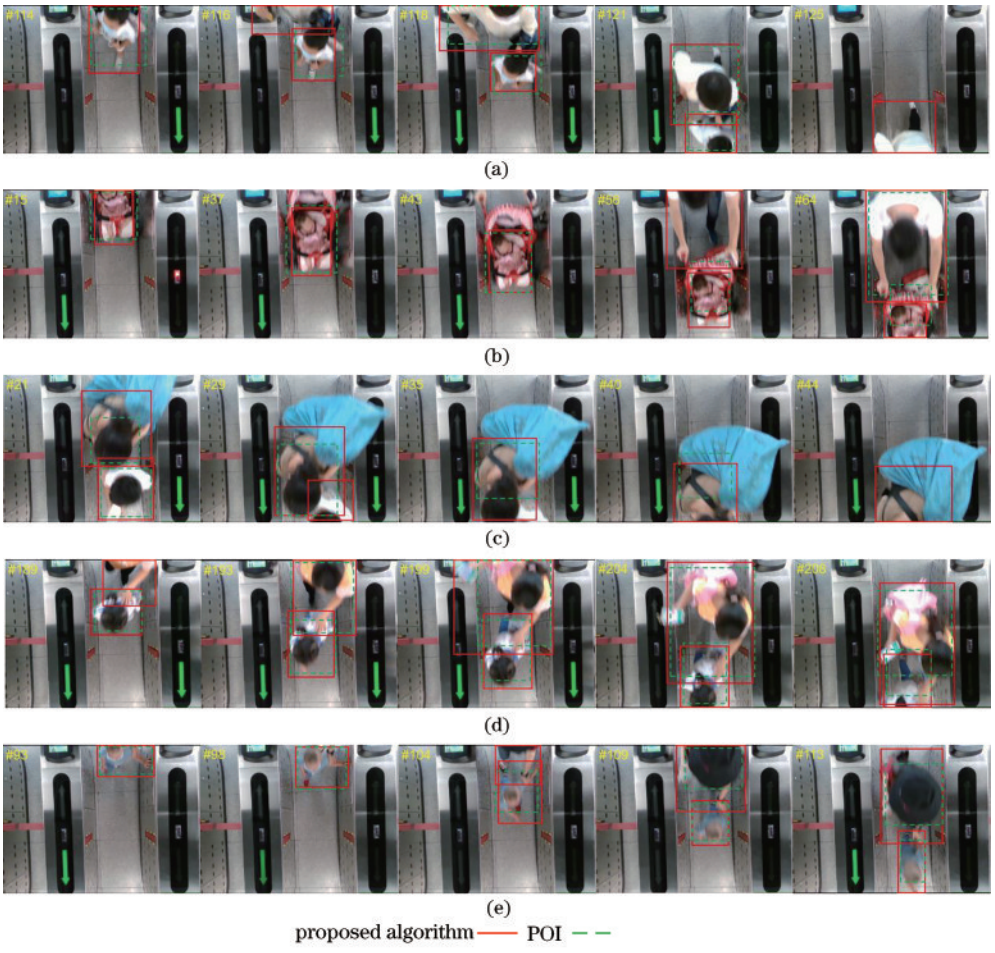


图 7　不同场景的跟踪结果对比。(a)进出视野；(b)背景干扰；(c)部分遮挡；(d)尺度变化；(e)形变

Fig. 7　Comparison of tracking results in different scenarios. (a) Out-of-view; (b) background clutters; (c) partial occlusion; (d) scale variation; (e) deformation

## 4　结　　论

提出了一种融合 YOLO-V4 与改进 SiameseRPN 的多目标跟踪算法。该算法将目标检测模型与目标跟踪模型进行融合，利用 YOLO-V4 的优势弥补 SiameseRPN 的不足，然后通过融合残差连接设计了轻量级特征提取网络，提高特征表征性能，采用自适应背景初始化策略增强模板特征的判别能力，最后通过数据关联匹配实现多目标的跟踪。实验结果表明：所提改进 SiameseRPN 算法在 VOT2016 和 VOT2018 数据集上的 EAO 指标与 SiameseRPN 相比分别提升了 6.53 个百分点和 6.5 个百分点，在 OTB100 上 AUC 和 Prec 分别为 0.654、0.846；所提算法在自建乘客数据集上的 MOTP 为 0.891，MOTA 为 0.937，跟踪速度为 39 frame·s$^{-1}$，实现了实时跟踪，性能优于对比算法，在应对目标尺度、外观变化、部分遮挡等情况时能够稳定跟踪。后续可从降低目标特征维度方面展开研究，进一步提升跟踪实时性。

## 参　考　文　献

[1] Luo W H, Xing J L, Milan A, et al. Multiple object tracking: a literature review[J]. Artificial Intelligence, 2021, 293: 103448.

[2] Amri S, Barhoumi W, Zagrouba E. A robust framework for joint background/foreground segmentation of complex video scenes filmed with freely moving camera[J]. Multimedia Tools and Applications, 2010, 46(2/3): 175-205.

[3] Sun D Q, Roth S, Black M J. Secrets of optical flow estimation and their principles[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 2432-2439.

[4] Khare M, Srivastava R K, Khare A. Single change detection-based moving object segmentation by using Daubechies complex wavelet transform[J]. IET Image Processing, 2014, 8(6): 334-344.

[5] 岳鑫, 李亮亮, 王红军, 等. 拟合标定模型的火炮稳像标定方法研究[J]. 光学学报, 2021, 41(15): 1510001.
Yue X, Li L L, Wang H J, et al. Gun image stabilization calibration method based on fitting calibration model[J]. Acta Optica Sinica, 2021, 41(15): 1510001.

[6] Nummiaro K, Koller-Meier E, van Gool L. An adaptive color-based particle filter[J]. Image and Vision Computing, 2003, 21(1): 99-110.

[7] Du K, Ju Y F, Jin Y L, et al. Object tracking based on improved MeanShift and SIFT[C]//2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), April 21-23, 2012, Yichang, China. New York: IEEE Press, 2012: 2716-2719.

[8] Exner D, Bruns E, Kurz D, et al. Fast and robust CAMShift tracking[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 9-16.

[9] Li B, Yan J J, Wu W, et al. High performance visual tracking with Siamese region proposal network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8971-8980.

[10] 王殿伟, 方浩宇, 刘颖, 等. 一种基于改进SiameseRPN的全景视频目标跟踪算法[J]. 激光与光电子学进展, 2020, 57(24): 241008.
Wang D W, Fang H Y, Liu Y, et al. Algorithm for panoramic video tracking based on improved SiameseRPN [J]. Laser & Optoelectronics Progress, 2020, 57(24): 241008.

[11] 陈法领, 丁庆海, 罗海波, 等. 基于自适应多层卷积特征决策融合的目标跟踪[J]. 光学学报, 2020, 40(23): 2315002.
Chen F L, Ding Q H, Luo H B, et al. Target tracking based on adaptive multilayer convolutional feature decision fusion[J]. Acta Optica Sinica, 2020, 40(23): 2315002.

[12] 金立生, 华强, 郭柏苍, 等. 基于优化DeepSort的前方车辆多目标跟踪[J]. 浙江大学学报(工学版), 2021, 55(6): 1056-1064.
Jin L S, Hua Q, Guo B C, et al. Multi-target tracking of vehicles based on optimized DeepSort[J]. Journal of Zhejiang University (Engineering Science), 2021, 55(6): 1056-1064.

[13] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2021-05-04]. https://arxiv.org/abs/2004.10934.

[14] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2(1/2): 83-97.

[15] Yu J H, Jiang Y N, Wang Z Y, et al. UnitBox: an advanced object detection network[C]//Proceedings of the 24th ACM international conference on Multimedia, October 15-19, 2016, Amsterdam, The Netherlands. New York: ACM Press, 2016: 516-520.

[16] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[M]// Hua G, Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 850-865.

[17] Bertinetto L, Valmadre J, Golodetz S, et al. Staple: complementary learners for real-time tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 1401-1409.

[18] Danelljan M, Häger G, Khan F S, et al. Learning spatially regularized correlation filters for visual tracking [C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 4310-4318.

[19] Danelljan M, Häger G, Khan F S, et al. Convolutional features for correlation filter based visual tracking[C]// 2015 IEEE International Conference on Computer Vision Workshop, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 621-629.

[20] Li X, Ma C, Wu B Y, et al. Target-aware deep tracking [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 1369-1378.

[21] Guo Q, Feng W, Zhou C, et al. Learning dynamic Siamese network for visual object tracking[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1781-1789.

[22] Zhang Z P, Peng H W. Deeper and wider Siamese networks for real-time visual tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4586-4595.

[23] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5000-5008.

[24] Bewley A, Ge Z Y, Ott L, et al. Simple online and realtime tracking[C]//2016 IEEE International Conference on Image Processing, September 25-28, 2016, Phoenix, AZ, USA. New York: IEEE Press, 2016: 3464-3468.

[25] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing, September 17-20, 2017, Beijing, China. New York: IEEE Press, 2017: 3645-3649.

[26] Kim C, Li F X, Ciptadi A, et al. Multiple hypothesis tracking revisited[C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 4696-4704.

[27] Yu F W, Li W B, Li Q Q, et al. POI: multiple object tracking with high performance detection and appearance feature[M]//Hua G, Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 36-42.

[28] Lu Y Y, Lu C W, Tang C K. Online video object detection using association LSTM[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2363-2371.