Opto-Electronic Engineering
光 电 工 程

Article

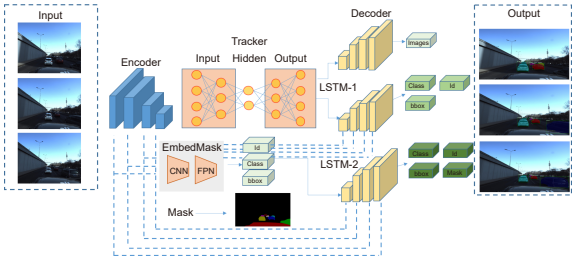
2022 年, 第 49 卷, 第 9 期

DOI: 10.12086/oee.2022.220024

融合空间掩膜预测与点云投影的多目标跟踪

陆康亮[1]，薛　俊[1]，陶重犇[1,2*]

[1]苏州科技大学电子与信息工程学院，江苏 苏州 215009；
[2]清华大学苏州汽车研究院，江苏 苏州 215134



摘要： 针对自动驾驶目标跟踪领域中，目标遮挡引起特征点损失，从而导致丢失跟踪目标的问题，本文提出了一种融合空间掩膜预测与点云投影的多目标跟踪算法，以减少遮挡产生的不利影响。首先，通过实例分割掩膜提取模型处理时序图像数据，获得基掩膜数据。其次，将获取掩膜数据输入跟踪器，通过预测模型获取后续序列图像掩膜输出，并利用验证器进行对比分析，以获得准确的目标跟踪输出。最后，将获取的二维目标跟踪数据投影到对应的点云图像中，获得最终的三维目标跟踪点云图像。本文在多个数据集上进行仿真实验，实验结果表明本文算法的跟踪效果优于其他同类算法。此外，在实际道路上进行测试，对于车辆的检测精度达到 81.63%，验证了本文算法也可以满足实际路况下目标跟踪的实时性需求。

关键词： 目标跟踪；空间掩膜预测；实例分割；点云投影

中图分类号： TP391.4　　**文献标志码：** A

陆康亮，薛俊，陶重犇. 融合空间掩膜预测与点云投影的多目标跟踪 [J]. 光电工程，2022，**49**(9): 220024

Lu K L, Xue J, Tao C B. Multi target tracking based on spatial mask prediction and point cloud projection[J]. *Opto-Electron Eng*, 2022, **49**(9): 220024

Multi target tracking based on spatial mask prediction and point cloud projection

Lu Kangliang[1], Xue Jun[1], Tao Chongben[1,2*]

[1]School of Electronics and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu 215009, China;
[2]Tsinghua University Suzhou Automotive Research Institute, Suzhou, Jiangsu 215134, China

Abstract: In the field of automatic driving target tracking, there is a problem that the target occlusion will cause the loss of feature points, resulting in the loss of tracking targets. In this paper, a multi-target tracking algorithm combining spatial mask prediction and point cloud projection is proposed to reduce the adverse effects of the occlusion. Firstly, the temporal image data is processed by an example segmentation mask extraction model, and the basic mask data is obtained. Secondly, the obtained mask data is input into the tracker, the mask output of subsequent sequence images is obtained through the prediction model, and the verifier is used for a comparative analysis to obtain an accurate target tracking output. Finally, the obtained 2D target tracking data is projected into

收稿日期：2022-03-22；收到修改稿日期：2022-06-06
基金项目：国家自然科学基金资助项目 (61801323，61972454)；中国博士后科学基金 (2021M691848)；苏州市科技项目 (SS2019029, SYG202142)
*通信作者：陶重犇，chongbentao@usts.edu.cn。

the corresponding point cloud image to obtain the final 3D target tracking point cloud image. In this paper, simulation experiments are carried out on multiple data sets. The experimental results show that the tracking effect of this algorithm is better than other similar algorithms. In addition, this paper is also tested on the actual road, and the vehicle detection accuracy reaches 81.63%. The results verify that the algorithm can also meet the real-time requirements of target tracking under the actual road conditions.

Keywords: target tracking; spatial mask prediction; instance segmentation; point cloud projection

1 引 言

在计算机视觉领域中，目标跟踪在自动驾驶中扮演着关键角色[1-2]。点云中的三维目标跟踪应用更是至关重要[2-4]。现有大多数三维目标跟踪算法都是直接处理点云信息[3]，然而，点云的稀疏性和无序性给这项任务带来了巨大的挑战，对于点云精确度的严重依赖也使得计算成本大幅度提高。本文将注意点从直接处理点云转移至利用二维数据处理三维数据。

随着视频处理能力取得巨大的进步，视频目标跟踪任务成为了新热点，国内外对目标跟踪进行了深入的研究[5-7]。在相关跟踪的方法中，如 Jiang 等[8] 提出的方法是一种基于注意力模型的分层多模式融合全卷积神经网络构建的 RGB-D 跟踪算法，能够通过深度图形获得更丰富的语义感知；Muller 等[9] 提出的方法从一系列 RGB-D 帧中检测每个帧中的目标并学习预测目标的几何形状以及到规范空间的密集对应映射，为每帧中的目标导出 6DoF 姿势和目标在帧之间的对应关系，从而在 RGB-D 序列中提供稳健的目标跟踪。上述算法都依赖于 RGB-D 数据，对于点云精度依赖性不高，因此计算成本低。此外，He 等[10] 提出了一种可学习图匹配方法的多目标跟踪，图匹配方法着重于轨迹和检测之间的关系。然而这些算法仍存在光照遮挡等问题导致的目标丢失情况，因此本文将注重优化跟踪任务中的目标丢失问题。

为了从数据中挖掘信息，国内外研究人员引入了自监督方法[11-12]。许多新的跟踪方法通过多模态融合来估计运动，例如 Luo 等[13] 通过利用来自点云的免监督信号和成对的相机图像来通过自监督纯粹地估计运动。研究人员还探索了利用视频的空间信息来进行学习的方法，如 Han 等[14] 通过对原始视频进行自监督的对比学习，强化视频目标的表示，从而进行动作识别；Wang 等[15] 采用自监督学习的方法来训练局部相关性模块，使得模型对相似物体的判别能力更强。然而，上述算法运算量大，并且无法满足目标跟踪所需的实时性要求。因此，本文将结合自监督优化方法设计一种跟踪器，并用于优化跟踪任务中的实时性。

针对上述问题，本文提出了一种基于融合空间掩膜预测与点云投影的多目标跟踪算法。算法模型主要在时序跟踪器前增加第一层掩膜输入以提高目标初始位置的准确性。其次，使用卷积神经网络为主导的跟踪器进行跟踪，优化每帧的单独实例分割，以提高跟踪器的速度，从而提高实时性效果。与之前的自监督方法类似，跟踪器对目标掩膜的空间特征学习，预测目标在后续帧的具体位置，从而获得整个视频序列对于目标的跟踪掩膜，因此减少遮挡对跟踪目标的丢失情况。还设置了一个验证层，将抽取样例掩膜输出与跟踪器对应图片输出进行比较验证，以减少检测不细致导致的跟踪丢失。最后，将二维视频处理后获得的掩膜数据，用点云投影的办法与三维雷达中的点云进行匹配，投影至三维点云中以获得三维目标跟踪。

创新点如下：1) 以卷积神经网络目标检测模块为基层模块，不直接使用实例分割模块进行跟踪，减少掩膜提取量。2) 增加了一个预测模块梯度损失函数，以增加算法对于预测掩膜精准度的把控，提高了算法对于预测出现差错所拥有的修正能力。3) 通过将二维跟踪到的目标掩膜数据投影到对应的点云图像上，不需要对于点云图像进行进一步处理。

2 理论推导

首先，本文利用空间位置预测模块来保存跟踪实例的位置信息，将序列视频信息输入算法跟踪层，以获得简单的跟踪目标信息。然后，通过基于 FCOS[16] 的 EmbedMask[17] 算法的掩膜分类分支来预测检测到的边界框，并且生成实例特定的空间系数。在空间掩膜预测模块中，针对边界盒内的每个子区域的空间进行单独的掩膜边界框预测，并通过每项单独的映射预测组合来获得最终实例掩膜预测。最后，通过确定信

息的实例掩膜数据与雷达点云的关联特征相对应，利用二维掩膜覆盖确定三维点云中的目标位置，算法框架图如图 1 所示。

2.1 掩膜预测模块

掩膜预测模块对同一物体不同帧的掩膜信息进行标定，从而获得每帧之间对于同一目标的跟踪定位。本模块对第一帧图像掩膜进行提取工作，后续帧的掩膜信息通过预测形式进行提取，这减少了提取图像每帧掩膜信息和跟踪对比所浪费的大量算力。同时，通过预测掩膜与验证帧掩膜的对比验证，这样既保证了掩膜预测正确，也节省了算力并提高运行速度，掩膜预测模型如图 2 所示。

2.1.1 基掩膜提取

目标检测的任务是找出图像中所有感兴趣的目标，并确定其位置和大小，这是机器视觉领域的核心问题之一[18]。随着深度学习的不断发展，实例分割算法开始出现，其目的是指定一个像素级的掩膜来对图像中的每个目标进行定位和分类。传统的目标检测仅提供盒级定位信息，然而实例分割提供目标轮廓级信息，从而能够更好地提供目标准确的位置信息。目标跟踪器需要对于目标捕捉拥有一定的实时性要求，然而 Mask RCNN[19] 算法存在无法保留精细掩膜和识别速度缓慢的弱点，难以做到实时准确的目标识别。因此，采用可高速运行并制作高分辨率掩膜的 EmbedMask 算法，这一优化能够快速地预处理图像序列，从而提高每秒输出帧数 (frames per second, fps)。此外，由于选取 EmbedMask 算法提供目标跟踪所需求的目标位置信息，单阶段方法减少了计算成本，从而能够给予后续跟踪器充足时间进行帧与帧对比分析以做到实时化处理。

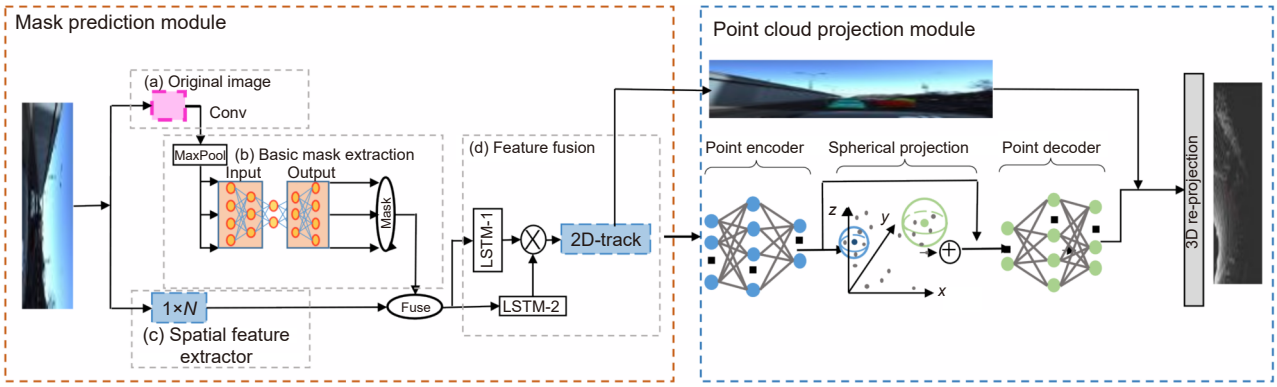


图 1 算法框架图

Fig. 1 Algorithm frame diagram

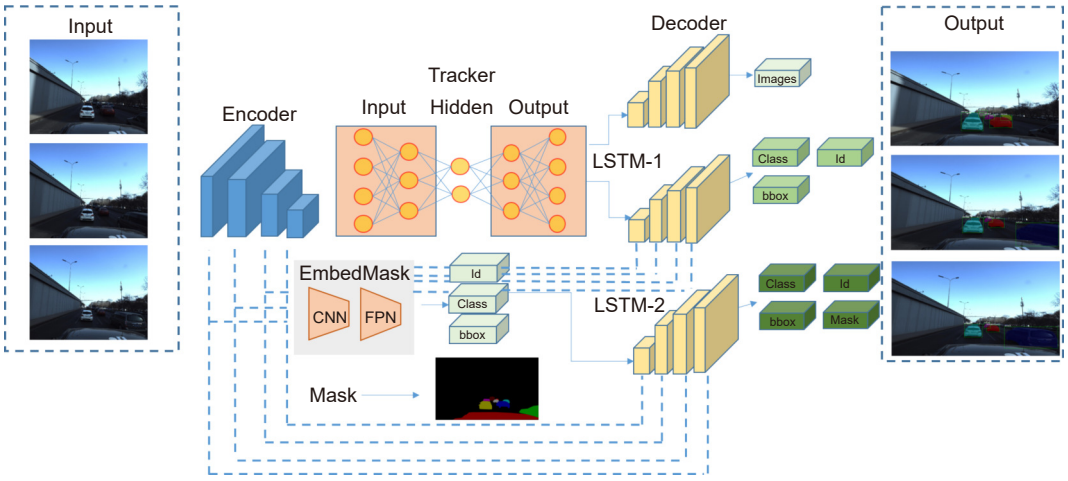


图 2 掩膜预测模块

Fig. 2 Mask prediction module

EmbedMask 算法是使用多任务丢失的端到端优化，其损失函数被联合优化为

$$L = L_{\text{cls}} + L_{\text{bbox}} + \lambda_1 L_{\text{mask}} + \lambda_2 L_{\text{match}} + L_{\text{center}}\,, \tag{1}$$

其中：除了 FCOS 中原有的分类损失L_{cls}和盒回归损失L_{bbox}之外，还引入了额外的损失用于掩膜预测的L_{mask}和L_{match}。

2.1.2 掩膜预测模块

掩膜预测模块对序列图像后续掩膜进行预测，使用一种特殊的循环神经网络 Long-Short Term Memory (LSTM)[20] 的门控制去预测获得最终的掩膜覆盖。LSTM 的特点在于单个循环结构内部有四个状态，循环结构之间将保持一个持久的单元状态不断传递下去，用于决定要遗忘或继续传递下去的信息。通过各种状态门之间的转化，判断该层是否选择输出或者跳转到下一层。通过阀门的控制，使得 LSTM 框架解决在训练过程中数据量过大而造成的梯度爆炸问题，这种算法很好地应用到序列图像中的目标数据关联计算。

当建立多目标跟踪外观模型时，可以认为c_t代表目标外观的存储模板，通过外观的数据变化量，决定是否将门o_t输出之前存储的外观c_{t-1}，以决定当前输出h_t。LSTM 使用以下规则运行：

$$c_t = f_t \circ c_{t-1} + i_t \circ g_t, h_t = o_t \circ \tanh(c_t)\,. \tag{2}$$

其中：$\circ$表示哈达玛积 (Hadamard product)。

基于 LSTM 的空间掩膜预测模块如图 3 所示。在给定输入的序列图像的情况下，预测模块以边界框、基掩膜和空间系数为输入，预测最终的掩膜。首先，通过基掩膜和空间系数定义模型中的外观输入c_t并存储第一帧的外观输入o_{t+1}，同时与第二帧的外观输入对比。然后，结合外观数据的变化量决定是否将门o_t输出预测外观o_{t+2}。最后，将新获得的外观输出和第三帧输入的外观数据进行比较获得一个新的L_{center}值以表达空间预测模块的梯度下降性能。

$$M_i = \sigma(B \times C_i)\,, \tag{3}$$

其中：σ为掩膜输入的规范化表示，B 是整个图像所拥有的 m 个基掩膜输入，C_i表示每张图片所获得的外观输入。M_i表示所有边界框的映射，已获得最终掩膜所拥有的边界框形状和位置信息。将预测边界框映射中心点与同帧边界框输入中心点信息对比，获得对于空间位置信息的变化梯度增加变量L_{center}如式 (4)，以确保边界框预测合理。

$$L_{\text{center}} = \nabla(M_i, M_{i+1})\,. \tag{4}$$

2.1.3 二维目标跟踪器

本文采用 EmbedMask 算法进行序列图像的边界框及基掩膜特征提取，只使用基掩膜和目标的边界框判定，减少了实例分割模块对于每一帧图像中每一个目标的掩膜分割。不使用每帧的掩膜覆盖输出作为跟踪器的输入，避免了输入数据量过大而导致的梯度爆炸问题，同时也减少了输入量递增所产生的检测速度极度下降的情况。其次，将输入的序列图像边界框判定以及基掩膜作为外观总特征，输入空间掩膜预测模型中，通过 LSTM 模型对于每帧图像外观特征的变化进行一定量约束，以门类型的决策进行输出判断。然后，对后续图像的掩膜输出进行预测，以获得最终的掩膜输出，从而实现多目标跟踪器。最后，通过验证帧的掩膜输入作为判据，以确保输出的掩膜帧跟踪目标的鲁棒性。通过这种优化减少一定量的特征输入并确保算法不失准确性，为提高实时速率提供了保障。

本算法采用实例分割检测模块与 LSTM 空间掩膜预测模块融合的多目标跟踪器。通过将不同帧的视频信息作为输入Σp_i，最终获得的掩膜Σy_i为输出，得出以下算法：

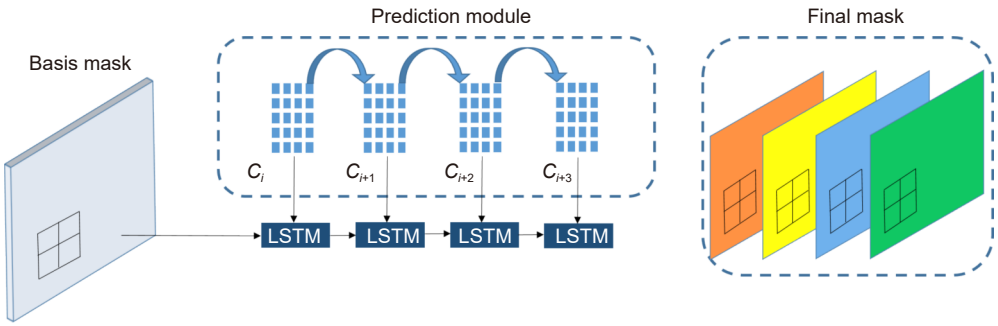


图 3 预测帧模块

Fig. 3 Prediction module

输入：视频序列图像$\{p_1, p_2, \ldots, p^*\}$
输出：估计目标边界框 $M_t=(x_t, y_t)$ 和最终掩膜 mask
将第一帧图像输入 p_1 实例分割模块
　　获取图像基掩膜 m_1
将图像 p_i 输入边界框模型
　　获取边界框 m_i
重复
$\{m_1, m_2, \ldots, m_i\}$和 p_1 定义为外观特征 C_i
将 C_i 输入 LSTM 模型中
如果 $\Delta C_i<k$
　　$C_i+1\rightarrow h_i$
　否则
　　$C_i\rightarrow h_i$
使用 H_i 更新 M_t 和最终掩膜
直至序列结束

2.2 点云投影

三维点云数据存在处理数据量大和处理难度高的问题，然而二维图像处理技术较为成熟，并且可以通过提升硬件性能来提高处理速度。因此，本文通过对二维相机数据与三维雷达点云数据的对比匹配，将二维图像的跟踪数据投影到三维点云中。点云投影的方法不仅能够让自动驾驶车辆对于道路情况拥有更好的信息感知能力，还可以减少点云计算并提高算法的实时性，点云投影模块如图 4 所示。

2.2.1 三维雷达与长焦相机的联合标定

激光雷达和单目相机之间的标定是传感器所必需的，需要获取相机与激光雷达外参，将点云三维坐标系下的点投影到相机三维坐标系下。还需要通过相机标定获得相机内参，将相机三维坐标系下的点投影到成像平面。通过调整标定板的角度可以将三维激光雷达数据与相机数据相转换，以确定雷达与相机内参统一，标定板效果如图 5 所示。

在联合标定时，会使用基于投影的方法估计激光雷达点云和相机图像平面中的一组目标特征。可以通过以下方式对应，找到激光雷达到相机的变换：

$$(R_C^{L^*}, T_C^{L^*}) = \arg\min_{(\boldsymbol{R},\boldsymbol{T})} \sum_i d_{\text{dist}}(P(H_L^C(X_i)), Y_i), \quad (5)$$

其中：X_i是雷达特征的 (均匀) 坐标，Y_i是摄像机特征的坐标，P 为“投影图”，H_L^C是雷达框架的带有旋转矩阵$R_C^{L^*}$和平移$T_C^{L^*}$的摄像机帧变换，d_{dist} 是距离或误差的度量。

在一个确定目标顶点的新方法中，令 P_r 表示目

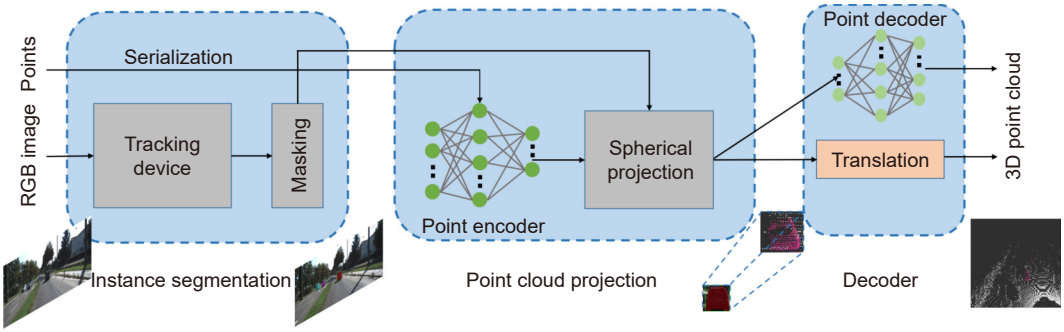


图 4　点云投影
Fig. 4　Point cloud projection

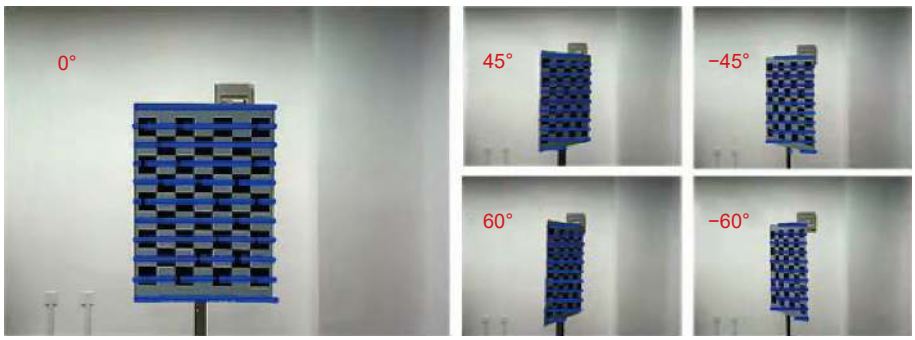


图 5　标定板效果
Fig. 5　Calibration board effect

标的雷达点云，并将三维点的集合设为X_i，使$P_r=\{X_i\}\,N_i=1$，其中N是目标上的点数。对于外部校准问题，需要估计雷达框架中的目标顶点。

对于$a>0$和$\lambda\in\mathbb{R}$有：

$$c(\lambda,a):=\begin{cases}\min(|\lambda-a|,|\lambda-a|) & \text{if } |\lambda|>a\\ 0 & \text{otherwise}\end{cases}. \tag{6}$$

令$H_{\mathrm{T}}^{\mathrm{L}}(P_r)=H_{\mathrm{L}}^{\mathrm{C}}(X_i)$，表示$H_{\mathrm{T}}^{\mathrm{L}}$对于点云的拉回，而$(\tilde{x}_i,\tilde{y}_i,\tilde{z}_i)$表示点的笛卡尔坐标，定义为

$$C\left(H_{\mathrm{T}}^{\mathrm{L}}(P_r):=\sum_{i=1}^{N}c(\tilde{x}_i,\epsilon)+c\left(\tilde{y}_i,\frac{d}{2}\right)+c\left(z_i,\frac{d}{2}\right)\right), \tag{7}$$

其中：d由 (正方形) 目标的大小确定，唯一的调整参数$\epsilon>0$(即理想目标的厚度)。并定义估计的目标顶点为

$$X_i^*:=H_{\mathrm{L}}^{\mathrm{T}*}(\bar{x}_i)\,,\ 1\leqslant i\leqslant 4\,. \tag{8}$$

首先通过三维雷达和相机的联合标定，获取了估计的目标顶点。有了对应关系后，下一步就是将雷达目标的顶点$[\ x_i\quad y_i\quad z_i\quad 1\ ]^{\mathrm{T}}=X_i$匹配到图像坐标中。

$$\begin{bmatrix}u'\\ v'\\ w'\end{bmatrix}=\begin{bmatrix}f_x & s & c_x\\ 0 & f_y & c_y\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1_{3\times3}\\ 0_{1\times3}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}R_{\mathrm{L}}^{\mathrm{C}} & T_{\mathrm{L}}^{\mathrm{C}}\\ 0_{1\times3} & 0\end{bmatrix}\begin{bmatrix}x_i\\ y_i\\ z_i\\ 1\end{bmatrix}, \tag{9}$$

$$L_i^Y=[\ u\quad v\quad w\ ]^{\mathrm{T}}=\begin{bmatrix}\dfrac{u'}{w'} & \dfrac{v'}{w'} & 1\end{bmatrix}^{\mathrm{T}}, \tag{10}$$

其中：式 (9) 包括相机的内部参数和外部参数$(R_{\mathrm{C}}^{\mathrm{L}},T_{\mathrm{C}}^{\mathrm{L}})$。从雷达坐标到图像坐标的匹配关系。

2.2.2 点云投影算法

假设已确定目标在雷达和相机图像平面中的顶点及其对应关系，用最小化相应角的欧几里得距离来表示外在变换的优化。同时，将对应投影多边形的并集和交集最大化。其中$C_i^Y\in\mathbb{R}^2$是摄像机角点公式

$$\begin{aligned}(R_{\mathrm{L}}^{\mathrm{C}*},T_{\mathrm{L}}^{\mathrm{C}*}):&=\arg\min_{\boldsymbol{R},\boldsymbol{T}}\sum_{i=1}^{4n}\left\|\prod(X_i;R_{\mathrm{L}}^{\mathrm{C}},T_{\mathrm{L}}^{\mathrm{C}})-C_i^Y\right\|_2^2\\ &=\arg\min_{\boldsymbol{R},\boldsymbol{T}}\sum_{i=1}^{4n}\left\|C_i^Y\right\|_2^2.\end{aligned} \tag{11}$$

最后，通过“交叉验证”的形式评估了此对应关系式。以循环方式使用一个或多个场景中的数据估算外在变换，然后评估剩下的场景。

当获得相机与雷达点云的对应关系后，将识别所得掩膜数据投影到雷达点云之中，通过球投影算法，将雷达中目标所对应点云数据进行渲染标注。以投影雷达点云的方式获得目标在三维点云图像中位置信息，投影关系如图 6 所示。

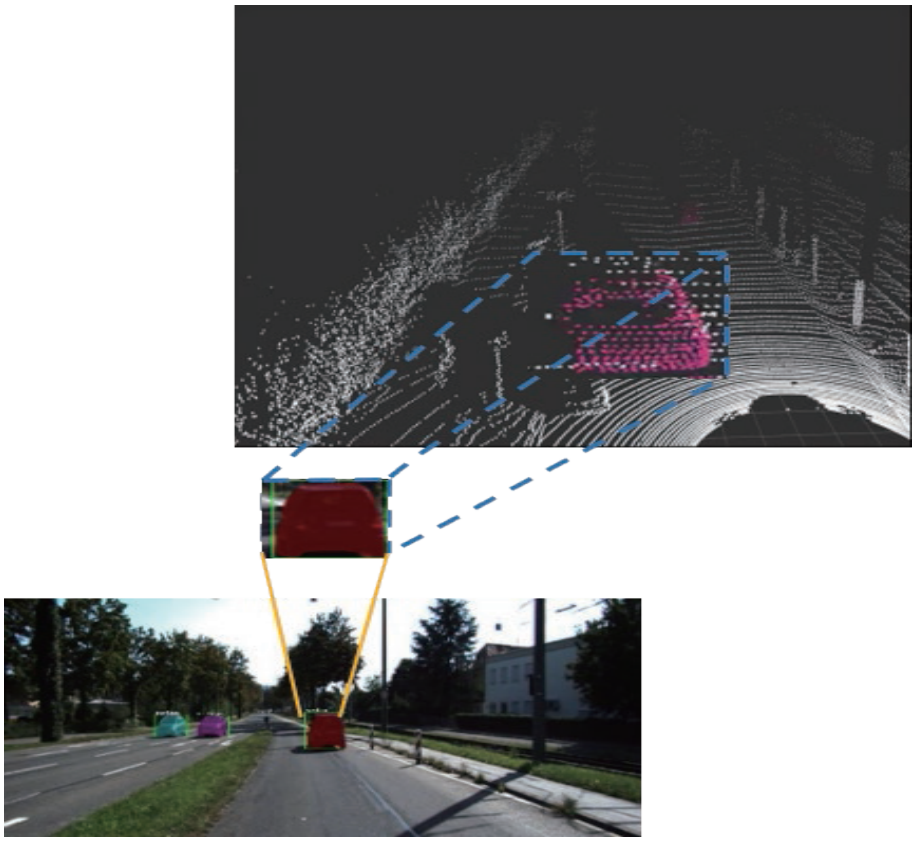


图 6 掩膜投影到点云中

Fig. 6 Mask projection

3 实验与分析

本文在阿波罗数据集、KITTI 数据集和 BDD100K 数据集中分别进行了实验。所选数据集中多组连续的道路时序图像存在大量车辆遮挡情况，能够很好地验证算法在自动驾驶过程中目标跟踪的效果。

3.1 二维跟踪器训练及分析

本文基于百度阿波罗数据进行实验模型训练。在训练模型中输入如图 7 的一组序列的视频数据和相对应的掩膜数据进行训练，最终获得跟踪器的权重文件。将阿波罗数据集的实例分割样本分为进行了训练、测试和验证三部分，对训练集进行了培训，并且在测试集和验证集中进行了训练结果的分析，再对边界框使用空间掩膜预测训练线性组合获得最终掩码，从而完善跟踪器。

在训练模型中输入所需数据，通过不断迭代和训练获得 loss 曲线如图 8 所示。其中图 8(a) 显示了算法定义的四个不同 loss 值所产生的曲线下降，图 8(b) 显示了四个 loss 值总和所展现的曲线梯度下降。通过曲线可以看出梯度下降在一开始表现得非常迅速，到后期逐渐趋于平缓。本算法通过四个 loss 值展示训练效果，最后将其总结成最终的 loss 值。

在掩膜提取部分，本文算法与主流实例分割算法进行对比，结果如表 1 所示。可以看出本文掩膜提取算法在识别精确度方面和主流的算法 Mask R-CNN 相近，并且实时速度提升显著，与主流算法 YOLACT 相近。通过上述实验分析，可以体现出本算法在保证精度稳定情况下，提高了算法的实时性。

利用 PR 曲线验证分类状况，并与 Yolo-V3[22]、Faster-RCNN[23]，MS-RCNN[24] 进行比较，PR 曲线如图 9 所示。其中 Precision 能够体现模型分类为正样本的数量中分类正确的比例，其计算方法为在设定某

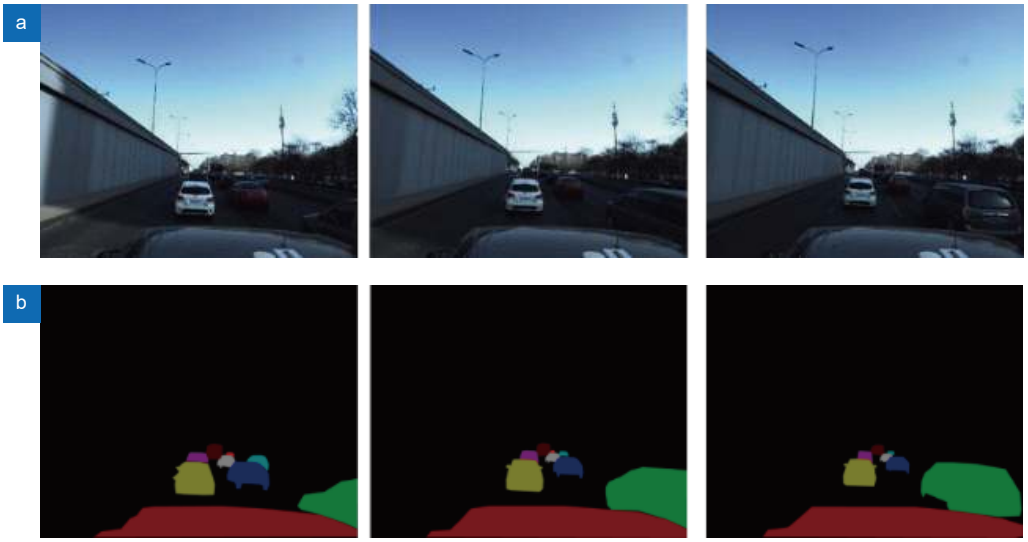


图 7 向模型中输入序列 RGB 和掩膜数据。(a) 原始序列 RGB 数据；(b) 对应序列掩膜数据

Fig. 7 Input sequence RGB and mask data into the model. (a) Original sequence RGB data; (b) Corresponding sequence mask data

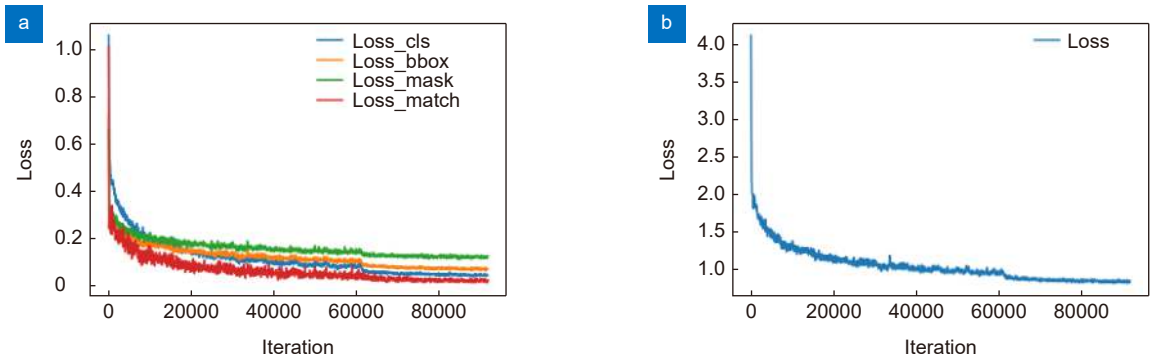


图 8 损失函数曲线。(a) 四种定义损失；(b) 总算法损失

Fig. 8 Loss function curve. (a) Four definitions of loss; (b) Total algorithm loss

表 1 本文与其他算法对比

Table 1 This algorithm is compared with other algorithms

Method	Backbone	ms	rc	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP^{bb}	fps
Mask R-CNN [19]	R-50-FPN			12	34.6	56.5	36.6	15.3	36.3	49.7	38.0	8.6
Mask R-CNN	R-101-FPN			12	36.2	58.6	38.5	16.4	38.4	52.0	40.1	8.1
Mask R-CNN	R-101-FPN	√		36	**38.1**	**60.9**	**40.7**	**18.4**	40.2	53.4	42.6	8.7
YOLACT-700 [21]	R-101-FPN	√	√	48	31.2	50.6	32.8	12.1	33.3	47.1	-	**23.6**
Ours	R-50-FPN			12	33.6	54.5	35.4	15.1	35.9	47.3	38.2	16.7
Ours	R-101-FPN	√		36	**37.7**	**59.1**	**40.3**	17.9	40.4	53.0	**42.5**	13.7
Ours-600	R-101-FPN	√		36	35.2	55.9	37.3	12.4	37.3	**54.9**	40.2	**21.7**

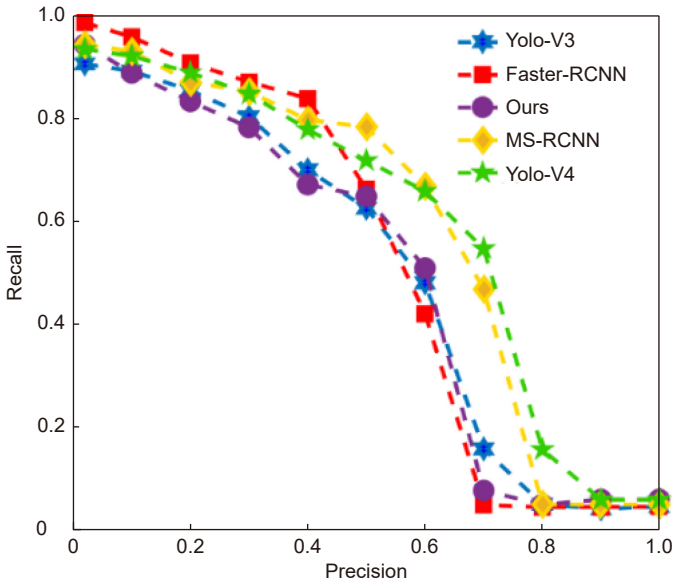


图 9 PR 曲线对比

Fig. 9 PR curve comparison

一阈值的情况下，正样本的预测数除以被预测为正样本的数量 (包含错误预测为正样本的负样本)。Recall 作为召回率，计算方法为在设定某一阈值的情况下，分类正确的样本除以所有正样本的数量。因此，不同的阈值会得到不同的 Precision 和 Recall 的值，由此绘制 PR 曲线。可以看出，本算法相较于这些成熟的目标检测算法具有良好的收敛性。在不同阈值下，本方法能够更好地兼顾精确率和召回率，且比其他方法收敛地更慢。

本文主要解决的是遮挡以及光度变化所导致的跟踪目标丢失而引起的跟踪器失灵的问题，因此本算法与其他算法在不同的距离、遮挡度、光度以及模糊度下进行对比实验。结果如图 10 所示，从图像看出，对于完全遮挡和完全图像损失，所有算法都完全丢失目标无法获得跟踪。然而，对于图像一般变化时，本文跟踪器的精确度明显高于另外两种跟踪器。特别对于遮挡情况，本算法能够在完全遮挡情况下，通过预测帧与帧的关联提供目标跟踪功能。这可以体现出本算法的鲁棒性强，对于遮挡问题的解决较为明显。

本文算法在传统的目标跟踪框架基础上，优化了遮挡导致的目标丢失问题。与其他算法在 MOT 数据集检测，结果如图 11 所示，在图中可以看出本算法的跟踪性能不输于其他传统方法，尤其在 KITTI 数据集中，本算法效果明显优于其他算法。

本算法与一些主流的多目标跟踪算法做了对比，结果如表 2 所示。在针对车辆的 KITTI 多目标跟踪数据集中，本算法的分割准确度优于其他所示算法。在针对行人的 KITTI 多目标跟踪数据集中，本算法效果与其他算法相近。由于本算法主要针对目标车辆进行跟踪，行人的外形变化可能会导致一定丢失目标，使得准确度略低于现有多目标跟踪算法。由于本算法采用空间掩摸预测模型对目标进行跟踪，在跟踪精确度方面略低于其他算法，然而本文的主要目标是提高算法对被遮挡目标的跟踪准确度，略微的跟踪精度下降

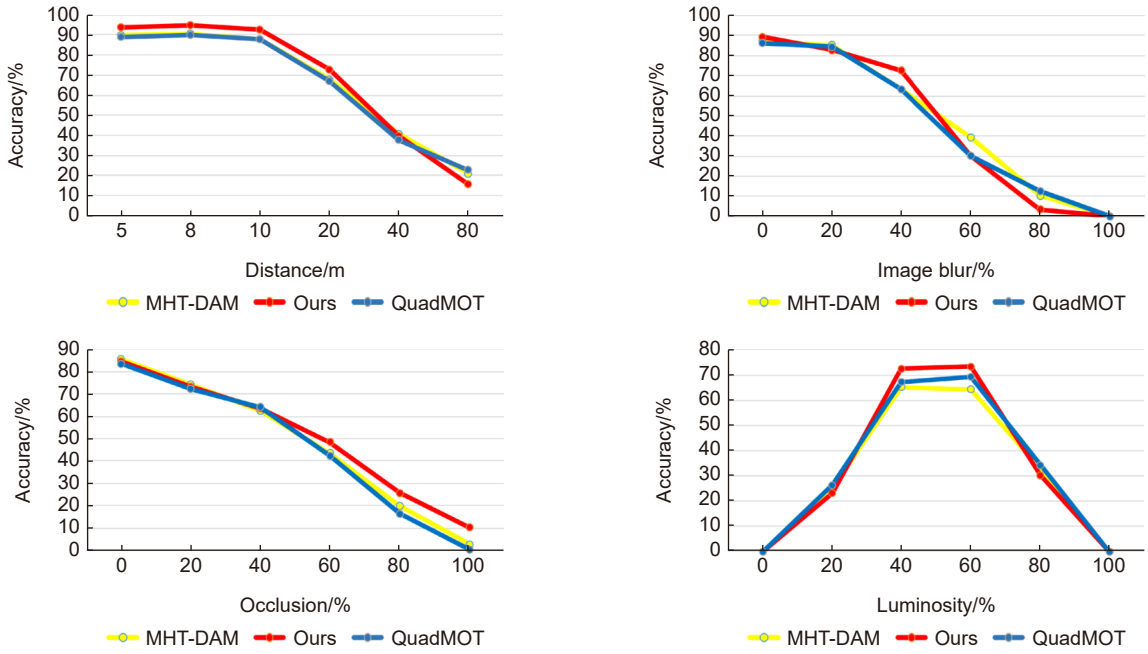


图 10 三种算法对于不同距离、遮挡度、光度以及模糊度的精确度

Fig. 10 The accuracy of the three algorithms for different distance, occlusion, luminosity and ambiguity

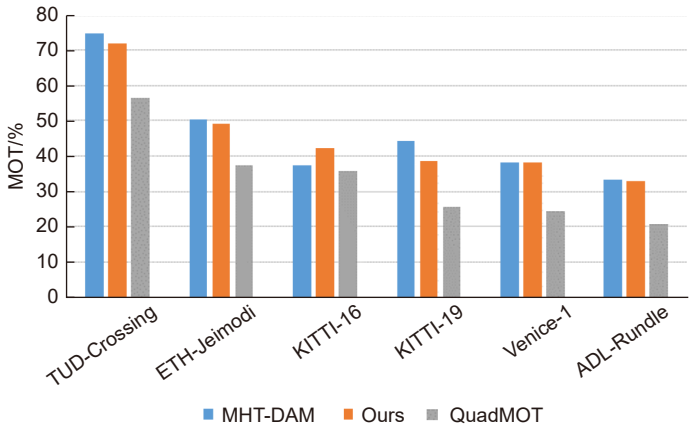


图 11 多种目标跟踪算法在 MOT 上的性能

Fig. 11 Performance of multiple target tracking algorithms on MOT

表 2 性能指标

Table 2 Performance index

	Method	sMOTSA	MOTSA	MOTSP
KITTI mots dataset-cars	Mask R-CNN	74.9	85.8	85.1
	MaskTrackR-CNN [25]	75.5	86.1	86.5
	Track R-CNN [26]	76.2	86.8	**87.2**
	Ours	**77.6**	**87.8**	86.3
KITTI mots dataset-pedestrains	Mask R-CNN	44.6	63.8	74.1
	MaskTrack R-CNN	45.9	64.6	**77.9**
	Track R-CNN	**46.8**	65.1	75.7
	Ours	45.3	**65.6**	77.0

不影响跟踪器效果体现。

在阿波罗数据集中测试结果如图 12 所示，对于目标的重叠等问题显示了良好的解决效果，所获得的跟踪目标准确并且掩膜覆盖得完整良好。

为了验证训练模型的鲁棒性，使用阿波罗数据中训练所获得的模型去测试 KITTI 数据集。如图 13 所示，对于不同情况的道路车辆状况，本算法具有一定的兼容性，这也为实时性开发奠定了良好的效果基础。对于 KITTI 数据集图像分辨率问题，采用了数据集匹配模型的方法进行了测试，后续将优化跟踪器模型对于不同相机的兼容性，以获得更好的实际使用效果。

在 BDD100K 数据集的多种经典无人驾驶子数据集中进行效果测试，可以更好地检测出算法对于复杂多变的交通环境的跟踪适应能力。在图 14 的表现中可以看出，本算法兼容性良好，对于不同的相机拍摄

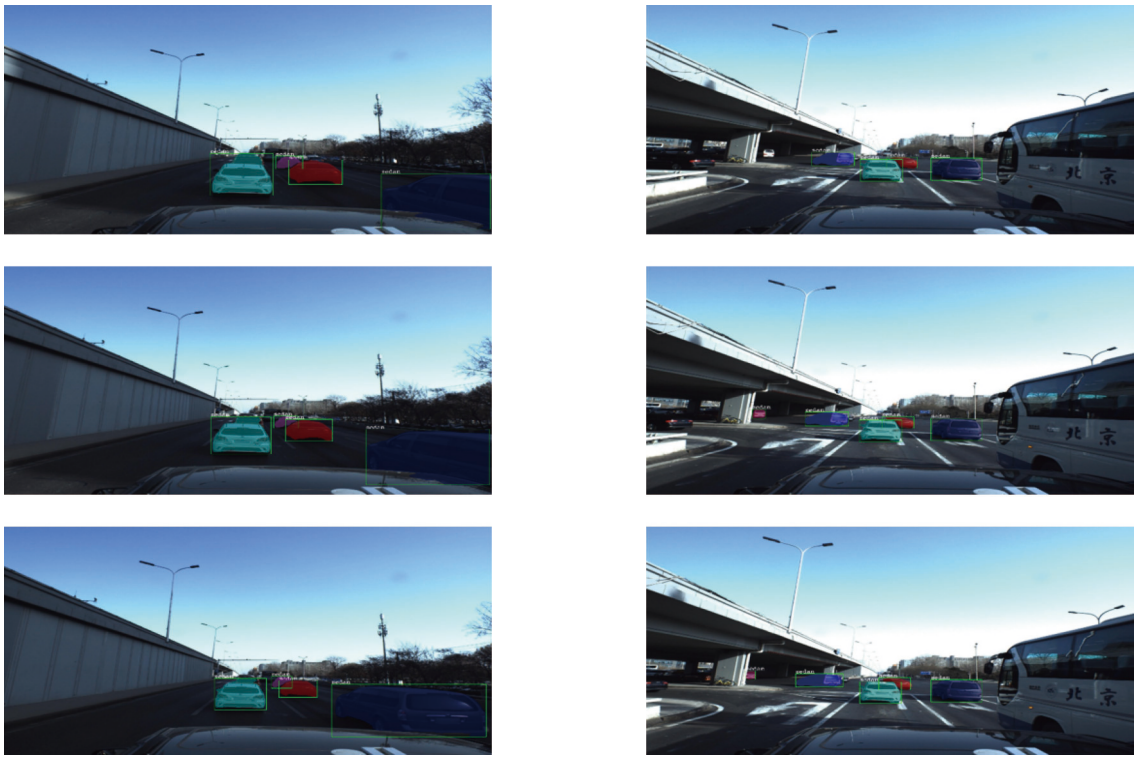


图 12　阿波罗数据集测试效果

Fig. 12　Effect of Apollo dataset test

图 13　KITTI 数据集测试效果

Fig. 13　Effect of KITTI dataset test

环境因素都拥有很好的稳定性，能够准确地检测出所要求的多目标跟踪车辆。

3.2 三维点云实验

本文通过联合标定的方法，将每帧获取的 RGB 目标跟踪掩膜与三维点云相匹配。以二维跟踪掩膜获取三维跟踪目标点云，从而获得三维目标跟踪图像。通过深度位置关系，可以使得自动驾驶车辆获得最为准确的目标信息，从而做出更好的决策。三维点云图像如图 15 所示，可以看到算法拥有较为良好的匹配。

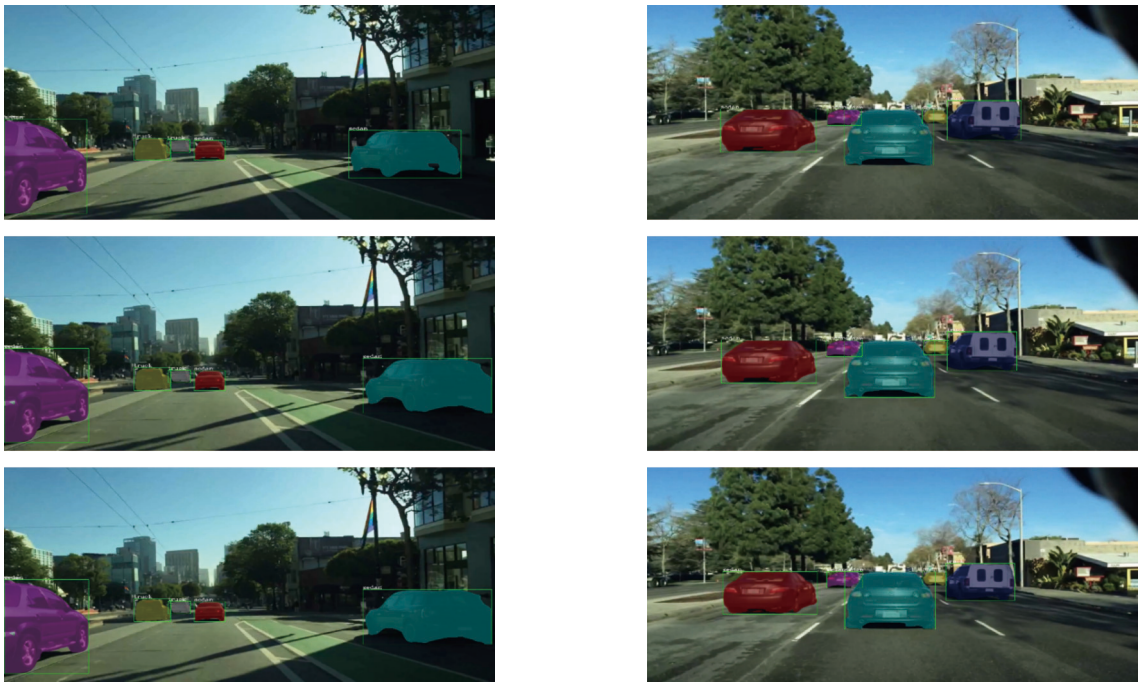


图 14 BDD100K 数据集测试效果

Fig. 14 Effect of BDD100K dataset test

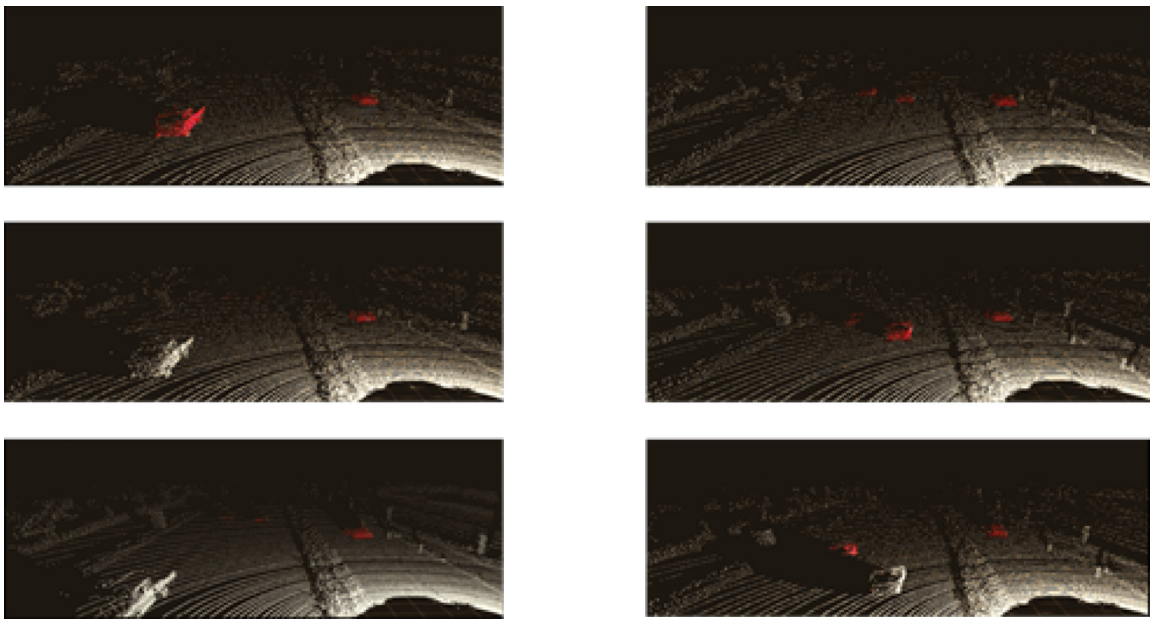


图 15 点云投影效果

Fig. 15 Effect of point cloud projection

3.3 道路测试

本文使用一辆搭载了多传感器的自动驾驶汽车进行实验。实验平台如图 16，汽车拥有一个集合了短焦相机的 32 线三维激光雷达、一个长焦相机、两个 RTK 定位装置和车头的毫米波雷达。主要使用三维激光雷达以及长焦相机进行配合，对点云目标进行跟踪处理。

实际道路测试在清华大学苏州汽车研究院的试验场地道路上进行，结果如图 17 所示。首先在 ROS 系统中录制实时道路数据，然后使用本算法对数据进行目标跟踪，最后可视化输出最终结果。本文构建的数据集输入模型得到的检测结果显示，整个算法在实验

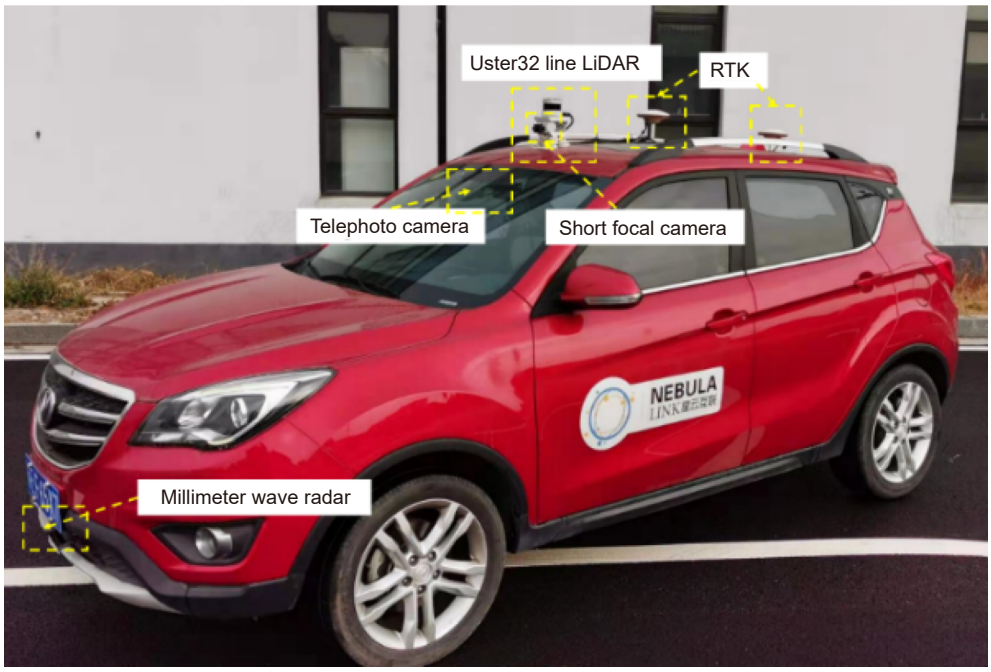


图 16 实验平台

Fig. 16 Experimental platform

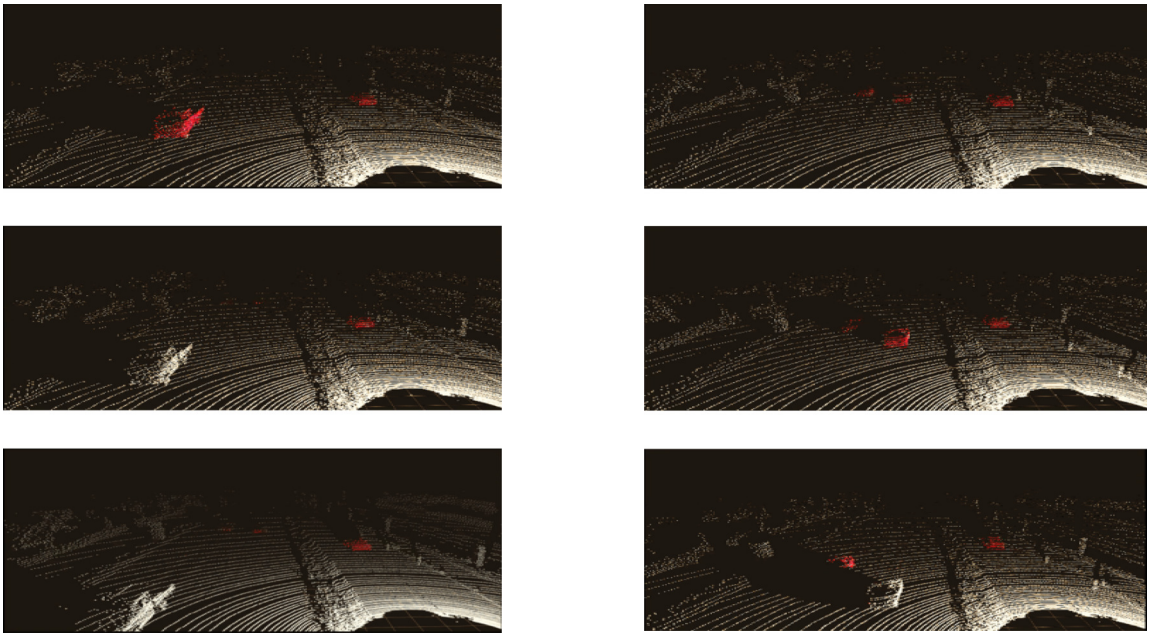


图 17 实际道路测试效果

Fig. 17 Effect of actual road experiment

平台上运行时间为 32 ms，对于车辆跟踪的平均精度达到 81.63%。从可视化结果也可以看出，本算法对中远距离的车辆检测效果良好，而对于近处的车辆，由于角度偏差较大存在丢失情况，需要进一步修正。上述实验表明，该算法计算效率高，可以满足实际路况下的实时检测需求。

4 结 论

本文提出基于掩摸预测与点云投影融合的多目标跟踪算法，很好地解决了目标跟踪中车辆重叠导致的目标丢失问题。在目标跟踪算法基础上，融合了空间掩摸预测来优化算法对于目标的跟踪效果，同时增加点云投影算法，将获取二维数据投影到三维点云中，能够更好地获取所跟踪的三维目标信息。结果表明，算法很好地解决了车辆遮挡导致的目标丢失问题。本文在多种数据集中进行跟踪器验证，能够很好地对多目标车辆进行跟踪。同时，算法在实际道路上进行测试，有效地满足了实际路况下的实时目标跟踪需求。后续将进一步增高算法对多辆车、拐弯、变道及车的形状突然变化情况处理的实时性和精准度，并通过传感器的边缘计算以更好地辅助决策系统判断。

参考文献

[1] Marvasti-Zadeh S M, Cheng L, Ghanei-Yakhdan H, et al. Deep learning for visual tracking: a comprehensive survey[J]. *IEEE Trans Intell Transp Syst*, 2022, **23**(5): 3943−3968.

[2] Chiu H K, Li J, Ambruş R, et al. Probabilistic 3D multi-modal, multi-object tracking for autonomous driving[C]//*Proceedings of 2021 IEEE International Conference on Robotics and Automation*, 2021: 14227–14233.

[3] Wu H, Han W K, Wen C L, et al. 3D multi-object tracking in point clouds based on prediction confidence-guided data association[J]. *IEEE Trans Intell Transp Syst*, 2022, **23**(6): 5668−5677.

[4] Tao C B, He H T, Xu F L, et al. Stereo priori RCNN based car detection on point level for autonomous driving[J]. *Knowl Based Syst*, 2021, **229**: 107346.

[5] Lv C, Cheng D Q, Kou Q Q, et al. Target tracking algorithm based on YOLOv3 and ASMS[J]. *Opto-Electron Eng*, 2021, **48**(2): 200175.
吕晨, 程德强, 寇旗旗, 等. 基于YOLOv3和ASMS的目标跟踪算法[J]. 光电工程, 2021, **48**(2): 200175.

[6] E G, Wang Y X. Multi-candidate association online multi-target tracking based on R-FCN framework[J]. *Opto-Electron Eng*, 2020, **47**(1): 190136.
鄂贵, 王永雄. 基于R-FCN框架的多候选关联在线多目标跟踪[J]. 光电工程, 2020, **47**(1): 190136.

[7] Jin L S, Hua Q, Guo B C, et al. Multi-target tracking of vehicles based on optimized DeepSort[J]. *J Zhejiang Univ (Eng Sci)*, 2021, **55**(6): 1056−1064.
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.

[8] Jiang M X, Deng C, Shan J S, et al. Hierarchical multi-modal fusion FCN with attention model for RGB-D tracking[J]. *Inf Fusion*, 2019, **50**: 1−8.

[9] Muller N, Wong Y S, Mitra N J, et al. Seeing behind objects for 3D multi-object tracking in RGB-D sequences[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 6067–6076.

[10] He J W, Huang Z H, Wang N Y, et al. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 5295–5305.

[11] Zhan X H, Pan X G, Dai B, et al. Self-supervised scene de-occlusion[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 3783–3791.

[12] Yuan D, Chang X J, Huang P Y, et al. Self-supervised deep correlation tracking[J]. *IEEE Trans Image Proc*, 2021, **30**: 976−985.

[13] Luo C X, Yang X D, Yuille A. Self-supervised pillar motion learning for autonomous driving[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 3182–3191.

[14] Han T D, Xie W D, Zisserman A. Video representation learning by dense predictive coding[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop*, 2019: 1483–1492.

[15] Wang Q, Zheng Y, Pan P, et al. Multiple object tracking with correlation learning[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 3875–3885.

[16] Tian Z, Shen C H, Chen H, et al. FCOS: fully convolutional one-stage object detection[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 9626–9635.

[17] Ying H, Huang Z J, Liu S, et al. EmbedMask: embedding coupling for one-stage instance segmentation[Z]. arXiv: 1912.01954, 2019. https://doi.org/10.48550/arXiv.1912.01954.

[18] Cao C L, Tao C B, Li H Y, et al. Deep contour fragment matching algorithm for real-time instance segmentation[J]. *Opto-Electron Eng*, 2021, **48**(11): 210245.
曹春林, 陶重犇, 李华一, 等. 实时实例分割的深度轮廓段落匹配算法[J]. 光电工程, 2021, **48**(11): 210245.

[19] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 2980–2988.

[20] Kim C, Li F X, Rehg J M. Multi-object tracking with neural gating using bilinear LSTM[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 208–224.

[21] Bolya D, Zhou C, Xiao F Y, et al. YOLACT: real-time instance segmentation[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 9156–9165.

[22] Redmon J, Farhadi A. Yolov3: an incremental improvement[Z]. arXiv: 1804.02767, 2018. https://doi.org/10.48550/arXiv.1804.02767.

[23] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//*Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015: 91–99.

[24] Tu Z G, Cao J, Li Y K, et al. MSR-CNN: applying motion salient region based descriptors for action recognition[C]//*Proceedings of the 2016 23rd International Conference on Pattern Recognition*, 2016: 3524–3529.

[25] Yang L J, Fan Y C, Xu N. Video instance segmentation[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 5187–5196.

[26] Voigtlaender P, Krause M, Osep A, et al. MOTS: multi-object tracking and segmentation[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 7934–7943.

作者简介

陆康亮 (2000-)，男，主要研究机器视觉中的目标跟踪与目标检测。

E-mail: 1324787850@qq.com

【通信作者】陶重犇 (1985-)，男，2014 年于江南大学获得博士学位，现为清华大学苏州汽车研究院博士后，主要研究方向三维语义建图和自主导航。

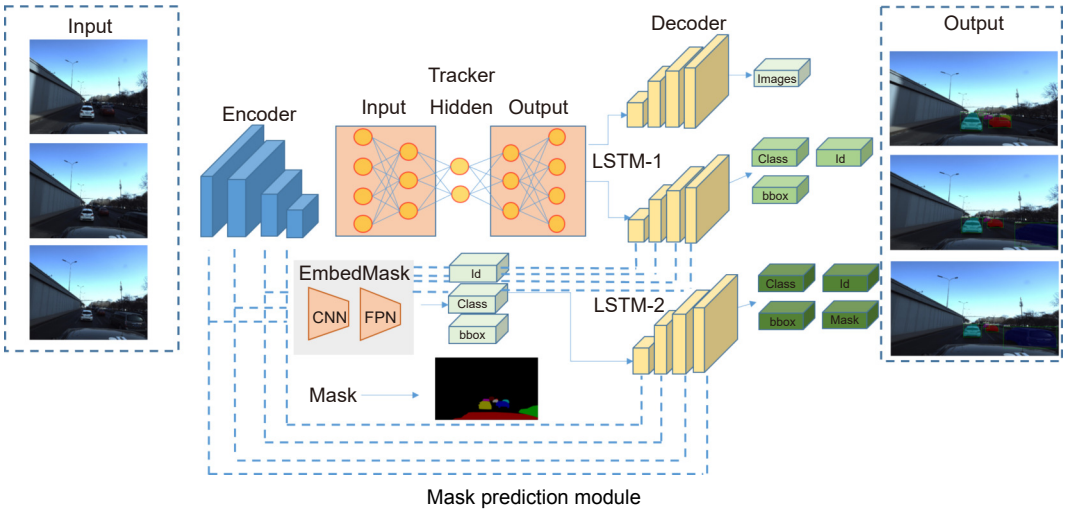
E-mail: chongbentao@usts.edu.cn

薛俊 (1998-)，男，硕士研究生，主要研究机器视觉中的目标跟踪与目标检测。

E-mail: 850765799@qq.com

Multi target tracking based on spatial mask prediction and point cloud projection

Lu Kangliang[1], Xue Jun[1], Tao Chongben[1,2*]



Overview: In the field of computer vision, target tracking plays a key role in automatic driving. In the field of automatic driving target tracking, there is the problem of feature point loss caused by the target occlusion, which will lead to the loss of tracking target. In this paper, a multi-target tracking algorithm combining spatial mask prediction and point cloud projection is proposed to reduce the adverse effects of the occlusion.

In this paper, the first mask input is added in front of the timing tracker to improve the accuracy of the initial target position. Secondly, the tracker dominated by a convolutional neural network is used for tracking, and the individual instance segmentation of each frame is optimized to increase the speed of the tracker, resulting in an improvement of the real-time effect. Finally, a verification layer is set to verify and compare the extracted sample mask output with the corresponding image output of the tracker, so as to reduce the tracking loss caused by the careless detection. After a two-dimensional video processing, the mask data is obtained, matched with the point cloud in the three-dimensional radar by the method of point cloud projection, and is projected into the three-dimensional point cloud to obtain the three-dimensional target tracking. The algorithm has three main contributions. Firstly, this paper takes the convolutional neural network target detection module as the basic module, does not directly use the example segmentation module for tracking, and extracts the mask of the first frame image. The mask information of the subsequent frame number can be extracted through the comparison and verification of the extraction of the prediction mask and the verification frame mask, which not only ensures the correct mask prediction, but also reduces the waste of a lot of computing power for extracting and tracking the mask information of each frame of the image. Secondly, this paper adds a gradient loss function of the prediction module to increase the control of the accuracy of the prediction mask and improve the correction ability of the algorithm for the prediction errors. Finally, this paper does not need to further process the point cloud image, but compares and matches the two-dimensional camera data with the three-dimensional point cloud radar data, and projects the two-dimensional tracked target mask data onto the corresponding point cloud image, so as to reduce the computational power of the algorithm and improve the real-time performance of the algorithm.

This paper is verified on the Apollo data set. The continuous road live screenshots are extracted from the data set to obtain the required set of time-series pictures, and the targets in the images are detected and tracked. Finally, the experiments show that the algorithm in this paper has an obvious effect on solving the occlusion problem. This paper has also been tested on the actual road, and the effect of medium and long-distance vehicle detection is good. The experiment shows that the algorithm can meet the real-time detection requirements under the actual road conditions.

Lu K L, Xue J, Tao C B. Multi target tracking based on spatial mask prediction and point cloud projection[J]. *Opto-Electron Eng*, 2022, **49**(9): 220024; DOI: 10.12086/oee.2022.220024

Foundation item: National Natural Science Foundation of China (61801323, 61972454), China Postdoctoral Science Foundation (2021M691848), and Science and Technology Projects Fund of Suzhou (SS2019029, SYG202142)

[1]School of Electronics and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu 215009, China; [2]Tsinghua University Suzhou Automotive Research Institute, Suzhou, Jiangsu 215134, China

* E-mail: chongbentao@usts.edu.cn