第 42 卷 第 16 期/2022 年 8 月/光学学报 研究论文

光学学报

基于特征融合和相似度估计网络的三维多目标跟踪

陈汶铭[1,2]，洪濡[1,2]，盖绍彦[1,2*]，达飞鹏[1,2,3**]

[1]东南大学自动化学院，江苏 南京 210096；

[2]东南大学复杂工程系统测量与控制教育部重点实验室，江苏 南京 210096；

[3]东南大学深圳研究院，广东 深圳 518063

摘要 针对现有自动驾驶多目标跟踪算法融合多传感信息的方式不能充分发挥协同作用的问题，提出了一种基于多模态特征融合与可学习式目标相似度估计的三维多目标跟踪算法。多模态特征融合模块对图像和点云特征进行基于通道注意力机制的特征融合，进一步提升了多模态特征的表达能力。目标相似度估计模块通过网络直接生成相似度矩阵，以可学习方式实现多目标之间的跨模态联合推理，避免了大量的人工参数设定。将所提算法在KITTI数据集上进行了验证与测试，其高阶跟踪精度(HOTA)在测试集中达到了69.24%，表明所提算法在精度上优于其他算法，具有较好的鲁棒性。

关键词 机器视觉；多目标跟踪；特征融合；注意力机制；卷积神经网络

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/AOS202242.1615001

Three-Dimensional Multi-Object Tracking Based on Feature Fusion and Similarity Estimation Network

Chen Wenming[1,2], Hong Ru[1,2], Gai Shaoyan[1,2*], Da Feipeng[1,2,3**]

[1]*School of Automation, Southeast University, Nanjing* 210096, *Jiangsu, China*;

[2]*Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing* 210096, *Jiangsu, China*;

[3]*Shenzhen Research Institute, Southeast University, Shenzhen* 518063, *Guangdong, China*

Abstract The multi-sensor information fusion method of the existing multi-object tracking algorithms for self-driving cannot give full play to synergy. To solve this problem, a three-dimensional multi-object tracking algorithm based on multi-modal feature fusion and learnable object similarity estimation is proposed. The multi-modal feature fusion module fuses the feature of images and point clouds on the basis of the channel attention mechanism to further improve the expressive ability of multi-modal features. The object similarity estimation module directly generates the similarity matrix through the network, and realizes the cross-modal joint reasoning between multiple objects in a learnable way, which avoids massive manual parameter setting. The proposed algorithm is verified and tested on the KITTI data set, and its higher-order tracking accuracy (HOTA) reaches 69.24% in the test set, which indicates that the algorithm is superior to other algorithms in accuracy and has good robustness.

Key words machine vision; multi-object tracking; feature fusion; attention mechanism; convolutional neural network

1 引 言

多目标跟踪是计算机视觉中非常重要的一个研究方向，在智能监控、自动驾驶等领域中有广泛的应用[1-3]。当前的多目标跟踪算法已经取得了很大的进展，但是在实际应用中会受到外部因素的制约如目标

收稿日期：2022-01-13；**修回日期**：2022-03-11；**录用日期**：2022-03-29

基金项目：国家自然科学基金(51475092)、江苏省前沿引领技术基础研究专项(BK20192004C)、深圳市科技创新委员会(JCYJ20180306174455080)

通信作者：*qxxymm@163.com；**dafp@seu.edu.cn

遮挡等，进而实现鲁棒的多目标跟踪仍然面临着严峻的挑战。

基于检测的跟踪范式分离了检测任务与跟踪任务，使得研究人员可以专注于跟踪任务的研究，因此受到了国内外学者的青睐。基于检测的跟踪范式依赖检测器提供的检测结果，跟踪任务负责给不同帧检测结果中的同一目标分配一致的身份(ID)[4]。早期的工作[5-7]使用的是目标的位置和尺寸信息，未能利用目标的外观和点云，虽然具有较高的效率，但是在目标遮挡等状况下，容易出现关联错误的情况。

Zhang等[8]将提取到的目标的二维(2D)特征和三维(3D)特征按维度相加作为融合特征。Vora等[9]先对图片进行语义分割，并对分割的结果对应的点云进行着色，再利用3D目标检测网络对点云进行处理。Shenoi等[10]将2D特征和3D特征同时输入到包含三层全连接层的网络中以计算得到新的融合特征，目标之间的相似度用融合特征的欧氏距离表示。Huang等[11]将图片和点云一起输入到区域建议网络中得到了融合特征。Chiu等[12]使用网络调整目标3D特征的维度，再将其与2D特征按维度相加。上述方法均使用了目标的外观特征和点云特征，按照融合方式可分为两类：1)使用较为复杂的流程与网络结构实现融合功能；2)使用按维度相加的融合方式。第一类方法增加了运算量，第二类方法要求特征的维度完全一致，并且相加过程中可能丢失一些信息。除此之外，高维度的特征包含了目标的语义信息，与欧氏距离、余弦距离等传统度量相比，利用深度神经网络的非线性学习能力设计的相似度估计网络较为合理。

针对上述问题，本文提出了一种基于多模态特征融合与可学习式目标相似度估计的3D多目标跟踪算法。多模态特征融合模块采用特征拼接作为基础融合方式，不要求不同模态特征维度完全一致，并且基于图像特征和点云特征，利用通道注意力机制调整各个通道的权重，强调重要信息，提升多模态特征的表达能力，提高跟踪精度。目标相似度估计模块根据特征计算相似度矩阵，相似度被映射到0～1之间，后续的数据关联不需要多次实验选择相似度阈值，避免了大量的人工参数设定工作。本文选择在KITTI数据集上验证所提算法性能，实验结果表明，所提算法能达到更高的精度，对外部环境有更强的鲁棒性。

2 所提算法

图1为所构建的多目标跟踪算法框架。第一步，将图片和点云作为输入，通过检测器获取图片和点云中目标的位置，如图1(a)所示。第二步，将图片中的目标裁剪出来，并通过VGGNet[13]提取目标的2D特征。同时，将点云中目标位置包含的点提取出来，并通过PointNet[14]提取目标的3D特征，如图1(b)所示。第三步，通过融合模块融合目标的2D和3D特征，获得目标的融合特征，如图1(c)所示。第四步，将融合特征输入进相似度估计网络中，生成相似度矩阵。原始相似度估计矩阵包含相邻前帧的所有检测与当前帧的所有检测间属于同一目标的得分和不属于同一目标的得分，如图1(d)所示。第五步，在数据关联部分，仅使用相似度矩阵中属于同一目标的得分并将之作为目标之间的相似度，使用贪心算法关联相似度高的目标，如图1(e)所示。在图1中，n为当前帧目标的数量，m为相邻前帧目标的数量，C为特征通道数。

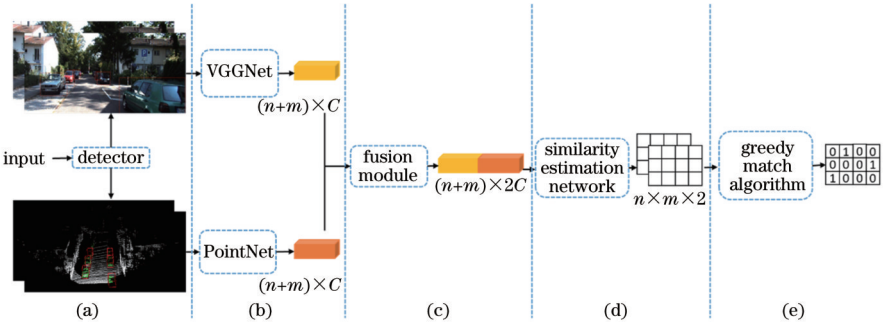

图1 所提算法流程图。(a)目标检测；(b)特征提取；(c)特征融合；(d)相似度估计；(e)数据关联

Fig. 1 Flow chart of proposed algorithm. (a) Object detection; (b) feature extraction; (c) feature fusion; (d) similarity estimation; (e) data association

2.1 特征提取

2D特征提取网络基于VGGNet11改进而来，输入是第t帧图片中的所有(n个)目标，每一个目标被裁剪下来并被变换为3×224×224维向量。完整的VGGNet为分类网络，本文为获取目标的2D特征，去掉原模型中的全连接层和Softmax层，并添加平均池化层以调整特征维度。2D特征提取网络模型如图2所示，其中Conv表示卷积层，ReLU表示线性整流函数。每个卷积层的步长为1，填充为1，仅改变特征的通道数，不改变大小。Maxpool表示最大池化层，步长为2，将特征的长和宽都缩小为原来的一半，不改变通道数。特征通过第5个最大池化层后，维度变为512×7×7。Avgpool表示平均池化层，将同一通道内的所有值求平均，最终输出的是512×1×1维特征向量。在基础数据不变的情况下，调整2D特征至512维。第t帧的2D特征可表示为$F_{2D}^{(t)}\in\mathbf{R}^{n\times 512}$，第$i$个目标的2D特征可表示为$f_{2D,i}^{(t)}\in\mathbf{R}^{1\times 512}$，$0\leqslant i<n$，其中$n$表示第$t$帧中目标的数量。

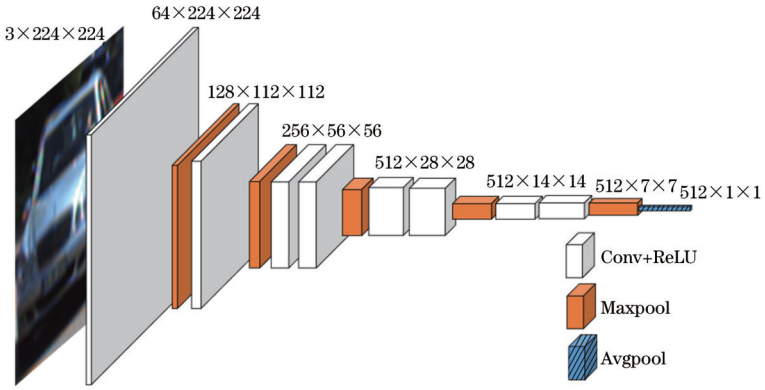

图 2　2D 特征提取网络模型

Fig. 2　2D feature extraction network model

本文的 3D 特征提取网络基于 PointNet 网络中分类部分进行改进，输入为所有表示目标的点，表示背景的点被去掉。原始 PointNet 网络的最终输出为目标在每个分类中的得分。本文为了获取点云中各个目标的 3D 特征，移除了负责分类的多层感知机（MLP），并增加了 Maxpool 层，并且以属于同一个目标的点特征作为输入，结构如图 3 所示。其中，MLP 后括号中的所跟数字表示 MLP 的输出维度，k 表示点云的数量。T-Net 网络学习一个转换矩阵，并将其与输入相乘以保证模型对特定空间变换的不变性。当特征被变换到 $k\times 512$ 维时，Maxpool 层将同属于一个目标的点的特征最大池化到 1×512 维，最终可得到 $n\times 512$ 维的 3D 特征。第 t 帧的 3D 特征表示为 $F_{3D}^{(t)}\in \mathbf{R}^{n\times 512}$，第 i 个目标的 3D 特征表示为 $f_{3D,i}^{(t)}\in \mathbf{R}^{1\times 512}$。

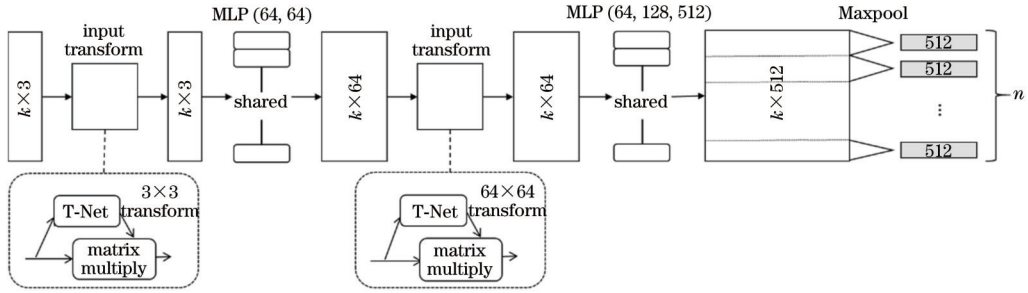

图 3　3D 特征提取网络模型

Fig. 3　3D feature extraction network model

2.2　特征融合

单模态特征有自身的局限性：仅使用图片信息，目标的外观相似会使得 2D 特征没有区分度，并且容易受到光照变化等外部环境的影响；仅使用点云信息，远处的点云过于稀疏会丢失目标。同时，由于特征各个维度对目标的贡献并不相同，为了强调更有表现力的特征，抑制不太重要的特征，因此引入注意力机制，使用挤压激励（SE）模块[15]调整特征权重。首先对原始特征中每个通道进行平均池化，得到 n 个特征描述符，这一步被称为挤压操作。在 n 维特征描述符向量经过全连接（FC）层、ReLU 层、全连接层和 Sigmoid 层后，将其作为权重与原始特征相乘，这一步被称为激励操作。最终，可得到权重被调整之后的特征。SE 模块可以描述为

$$F=f_{se}(F_{ori})=\sigma\left\{f_{FC}^{(C)}\left\{\delta\left\{f_{FC}^{(C/r)}\left[f_{avg}(F_{ori})\right]\right\}\right\}\right\}F_{ori},\quad(1)$$

式中：F_{ori} 为原始特征；$f_{se}(\cdot)$ 为 SE 模块；$f_{avg}(\cdot)$ 为通道内的平均池化；$f_{FC}^{(C/r)}(\cdot)$ 为输出维度为 C/r 的全连接层；C 为通道数；r 为缩放系数；$f_{FC}^{(C)}(\cdot)$ 为输出维度为 C 的全连接层；$\delta(\cdot)$ 为 ReLU 函数；$\sigma(\cdot)$ 为 Sigmoid 函数；F 为通过 SE 模块调整通道权重之后的特征。SE 模块如图 4 所示。

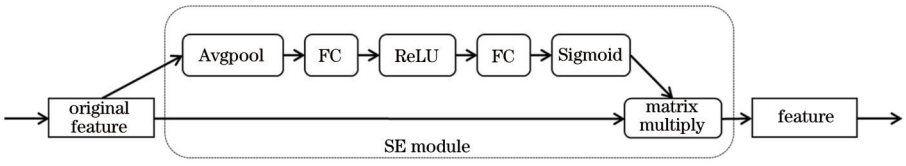

图 4　SE 模块

Fig. 4　SE module

为了让不同模态的特征弥补对方的缺陷，并且让各模态特征各维度都有表达的机会，本文以特征拼接为基础融合方法，并在此基础上引入注意力机制。先利用SE模块调整2D特征和3D特征，再拼接在一起作为融合特征，可以描述为

$$F_{\text{fuse}}^{(t)}=\text{CONCAT}\left\{f_{\text{se}}\left[F_{\text{2D}}^{(t)}\right],f_{\text{se}}\left[F_{\text{3D}}^{(t)}\right]\right\},\tag{2}$$

式中：CONCAT(•)表示特征拼接；$F_{\text{fuse}}^{(t)}\in\mathbf{R}^{n\times 1024}$；$F_{\text{2D}}^{(t)}$为原始2D特征；$F_{\text{3D}}^{(t)}$为原始3D特征。特征融合模块如图5所示。

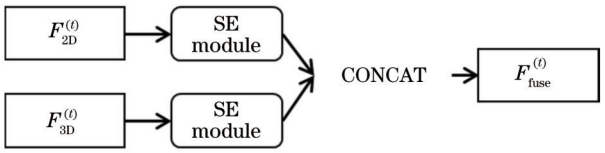

图5　特征融合模块

Fig. 5　Feature fusion module

先使用SE模块调整各个模态特征的通道权重，可以确保在特征拼接之前，各个模态特征中重要的通道权重已经被提高。

2.3　相似度估计

为了更有效地利用融合特征中的高维度信息，本文将第t帧中的任意目标和第$t-1$帧中的任意目标是否是同一目标，即是否ID一致看作是分类问题，并由此设计了二分类网络作为相似度估计网络。相似度估计网络的输入为第t帧中的n个目标的融合特征$F_{\text{fuse}}^{(t)}\in\mathbf{R}^{n\times 1024}$和第$t-1$帧中的$m$个目标的融合特征$F_{\text{fuse}}^{(t-1)}\in\mathbf{R}^{m\times 1024}$。第$t$帧和第$t-1$帧的特征通过复制扩展至$n\times m\times 1024$维，对应特征元素相减作为相似度估计网络的输入，可以描述为

$$F_{\text{match}}^{(t)}=F_{\text{fuse}}^{(t)}-F_{\text{fuse}}^{(t-1)},\tag{3}$$

式中：$F_{\text{match}}^{(t)}\in\mathbf{R}^{n\times m\times 1024}$。利用卷积层降低维度并进一步提取有用的特征，然后通过全连接层输出分类信息。令第t帧中的任意目标与第$t-1$帧中的任意目标的ID一致为正例，反之为负例，可以描述为

$$M=f_{\text{FC}}^{(2)}\left\{f_{\text{FC}}^{(20)}\left\{f_{\text{conv}}^{(1\times 1)}\left[F_{\text{match}}^{(t)}\right]\right\}\right\},\tag{4}$$

式中：$f_{\text{conv}}^{(1\times 1)}(\cdot)$表示核尺寸为$1\times 1$的卷积层；$f_{\text{FC}}^{(20)}(\cdot)$表示输出维度为20的全连接层；$f_{\text{FC}}^{(2)}(\cdot)$表示输出维度为2的全连接层；$M\in\mathbf{R}^{n\times m\times 2}$表示每一个目标对在两个分类上的得分。最后，利用Softmax函数将M中的数值映射到0～1之间，得到的相似度估计矩阵为$S\in\mathbf{R}^{n\times m\times 2}$。相似度估计网络结构如图6所示，其中Conv1-160表示核尺寸为1×1、输出维度为160的卷积层，FC后的数字表示全连接层的输出维度。

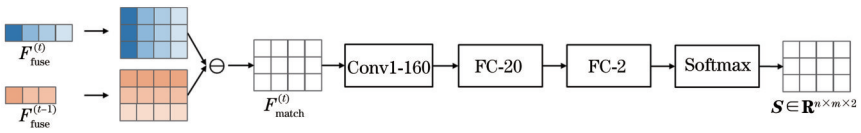

图6　相似度估计网络

Fig. 6　Structure of similarity estimation network

2.4　数据关联

数据关联模块接收相似度估计网络输出的相似度矩阵，取第一个通道，即$S\in\mathbf{R}^{n\times m}$，其中的每个数字代表对应目标ID一致的概率，$s_{i,j}=0.5$代表网络预测第$t$帧中第$i$个目标与第$t-1$帧中第$j$个目标ID一致的得分和ID不一致的得分相等。特别地，S中行代表第t帧中的目标，列代表第$t-1$帧中的目标。使用贪心算法关联数据，如图7所示。从第一行开始，选取得分最高的列，如果得分超过0.5，则认为此行与此列代表的目标是对应的，分配相同的ID，并且此列中的数据均被置为0，以防止被二次匹配。当得分低于0.5时，认为没有与此行代表的目标匹配的历史目标，为其分配新的ID。

3　实验与分析

3.1　数据集及评估方法

使用KITTI数据集[16]作为模型训练与测试的平台，使用高阶跟踪精度(HOTA)[17]作为评估指标。KITTI数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集之一，包含市区、乡村和高速公路等场景中采集的真实图像数据，每张图像中最多达15辆车和30个行人，还有各种程度的遮挡与截断，以10 Hz的频率采样和同步。KITTI数据集将车辆与行人分开评估，本文仅对车辆进行多目标跟踪。由于KITTI数据集中仅划分训练集与测试集，故本文自主选取官方训练集中的0001、0003、0004、0006、0008、0009、0012、0013、0015、0020号训练序列作为训练集，共计10个视频序列，包含3999帧图片和车辆的地面真值(ground truth)检测12957个。选取0000、0002、0005、0007、0010、0011、0014、0016、0017、0018、0019号训练序列作为验证集，共11个视频序列，包含4009帧图片和车辆的地面真值检测11113个。官方测试集中共29个视频序列，包含11095帧图片。进行消融实验与对比实验的各个网络仅在训练集上训练，并在验证集上验证，在测试集上测试。

HOTA是最新的评估跟踪效果的指标，它明确地平衡了检测、关联和定位的精度对于最终指标的影响并可以分解为与之相关的一系列子指标，包括关联精度(M_{AssA})、检测精度(M_{DetA})等。其中，关联精度侧重关联方面的得分，检测精度侧重检测方面的得分。

检测精度评估了所有的检测目标和所有真值之间

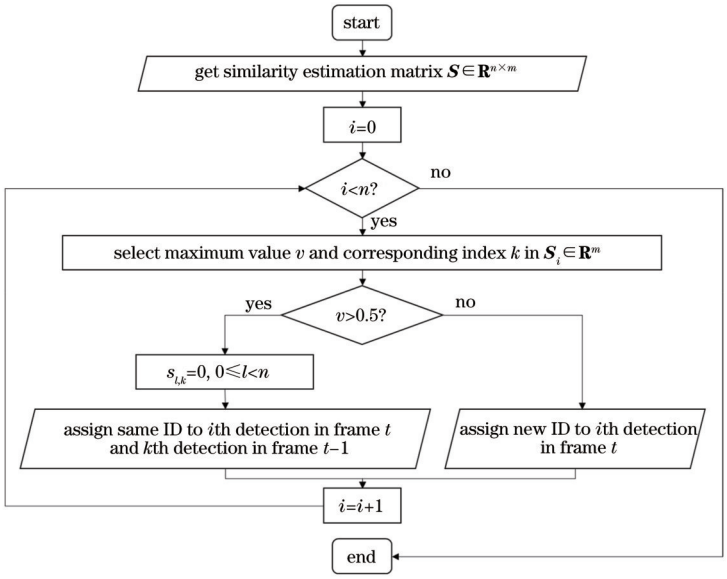

图 7 贪心匹配流程图

Fig. 7 Flow chart of greedy matching

对齐的情况，HOTA 使用检测交并比来计算检测精度，检测精度的表达式为

$$M_{\mathrm{DetA}}=\frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}}+n_{\mathrm{FN}}+n_{\mathrm{FP}}},\tag{5}$$

式中：下标 TP 表示真阳性检测，即一个检测与一个真值交并比高于某个设定的阈值，当多个检测与一个真值交并比高于阈值或者一个检测与多个真值交并比高于阈值时，使用匈牙利算法确定一对匹配；下标 FN 表示假阴性检测，即某真值未匹配到检测；下标 FP 表示假阳性检测，即某检测未匹配到真值；n_{TP}、n_{FN}、n_{FP} 为三种情况相应的数量。

关联精度评估了跟踪器将检测与时间关联起来的好坏程度，通过类似于交并比的方式可以测量这一点。关联交并比的表达式为

$$r_{\mathrm{ass}}=\frac{n_{\mathrm{TPA}}}{n_{\mathrm{TPA}}+n_{\mathrm{FNA}}+n_{\mathrm{FPA}}},\tag{6}$$

式中：下标 TPA 表示真阳性关联，即预测轨迹包含的检测与真实轨迹包含的真值之间能够相互匹配的部分；下标 FNA 表示假阴性关联，即真实轨迹中的真值在预测轨迹中没有对应的检测相匹配；下标 FPA 表示假阳性关联，即预测轨迹中包含的检测没有与之匹配的真值，或者此检测与其他真实轨迹中的真值相匹配；n_{TPA}、n_{FNA}、n_{FPA} 表示三种情况相应的数量。对整个数据集中的检测和真值计算的 r_{ass} 取平均来衡量关联精度，关联精度的表达式为

$$M_{\mathrm{AssA}}=\frac{1}{n_{\mathrm{TP}}}\sum_{c}r_{\mathrm{ass}}(c)=\frac{1}{n_{\mathrm{TP}}}\sum_{c}\frac{n_{\mathrm{TPA}}(c)}{n_{\mathrm{TPA}}(c)+n_{\mathrm{FNA}}(c)+n_{\mathrm{FPA}}(c)},\tag{7}$$

式中：c 为真阳性检测集合中的单个检测。

HOTA 指标包含了上述指标，可以作为单一的指标来衡量跟踪器的效果，其表达式为

$$M_{\mathrm{HOTA}}^{(\alpha)}=\sqrt{M_{\mathrm{DetA}}^{(\alpha)}M_{\mathrm{AssA}}^{(\alpha)}}=\sqrt{\frac{\sum_{c}r_{\mathrm{ass}}^{(\alpha)}(c)}{n_{\mathrm{TP}}^{(\alpha)}+n_{\mathrm{FN}}^{(\alpha)}+n_{\mathrm{FP}}^{(\alpha)}}},\tag{8}$$

$$M_{\mathrm{HOTA}}=\int_{0<\alpha\leqslant 1}M_{\mathrm{HOTA}}^{(\alpha)}\approx\frac{1}{19}\sum_{\alpha=0.05}^{0.95}A_{\mathrm{HOT}}^{(\alpha)}。\tag{9}$$

请注意，之前的 M_{DetA} 和 M_{AssA} 都是使用基于阈值 α 的匈牙利匹配来计算的，即检测与真值的交并比大于 α 时，视检测为 TP。取单个阈值下的 $M_{\mathrm{HOTA}}^{(\alpha)}$ 为 $M_{\mathrm{DetA}}^{(\alpha)}$ 和 $M_{\mathrm{AssA}}^{(\alpha)}$ 的几何平均值。然后，对不同的阈值进行积分，获得最终的 M_{HOTA}。M_{DetA}、M_{AssA} 和 M_{HOTA} 的数值越大，代表精度越高。

3.2 实验环境及学习参数

本文所用计算机的硬件配置：中央处理器（CPU）为 Intel® Core™ i5-10600KF，主频为 4.10 GHz；图形处理器（GPU）为 GeForce RTX 3070。深度学习框架使用 PyTorch 1.9.1，环境配置使用 anaconda 4.10.1。

优化方法为使用动量的随机梯度下降法，可以描述为

$$g=\frac{1}{m}\nabla_{\theta}\sum_{h=1}^{m}L\left\{f\left[x^{(h)};\theta\right],y^{(h)}\right\},\tag{10}$$

$$Mv-\varepsilon g\rightarrow v,\tag{11}$$

$$\theta+v\rightarrow\theta,\tag{12}$$

式中：g 表示从训练集中采样的 m 个样本的平均梯度；$L(\cdot)$ 表示交叉熵损失函数；$f(\cdot)$ 表示网络；θ 表示网络参数；$x^{(h)}$、$y^{(h)}$ 分别表示第 h 个输入和该输入相应的真值；v 表示速度；M 表示动量参数；ε 表示学习率。本文设置动量为 0.9，初始学习率为 0.005，训练轮数为 20 轮，每 5 轮学习率衰减一半。

3.3 细节补充

本文使用 DSA-PV-RCNN[18] 作为检测器，为跟踪器提供检测结果。为了降低假阳性检测对算法跟踪精度的影响，去除得分低于 0.3 的检测。

3.4 消融实验

3.4.1 特征融合模块

为了验证各个模态的特征对结果的贡献和探寻最合适的融合方法，在KITTI数据集上进行训练与验证。

本组实验设置了4组不同的模态：Image表示只使用图像信息，即只使用2D特征；Point表示只使用点云信息，即只使用3D特征；Fusion-add表示将2D特征和3D特征按维度相加后作为融合特征；Fusion-cat表示将2D特征和3D特征拼接起来作为融合特征。

同时，为了验证注意力机制的作用和不同注意力机制加入方式的效果，为不同的模态设置了几种方法。Image和Point是单模态网络，分别设置方法A和方法B，方法A使用原始的2D特征或者3D特征，方法B使用SE模块调整原始特征通道的权重，如图8所示。其中，$\tilde{F}_{2D}^{(t)}$和$\tilde{F}_{3D}^{(t)}$表示通过SE模块调整通道权重之后的特征。Fusion-add和Fusion-cat是多模态网络，分别设置融合方法C、融合方法D和融合方法E。其中，方法C表示将原始2D特征和3D特征直接相加或拼接，方法D表示将原始特征相加或拼接后通过SE模块调整通道权重，方法E表示将原始特征通过SE模块调整通道权重后进行相加或者拼接操作，如图9所示。其中，ADD表示特征相加操作，本文所提算法为特征拼接网络中的方法E。

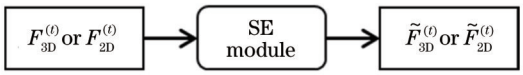

图8 单模态网络中的方法B

Fig. 8 Method B in single modal network

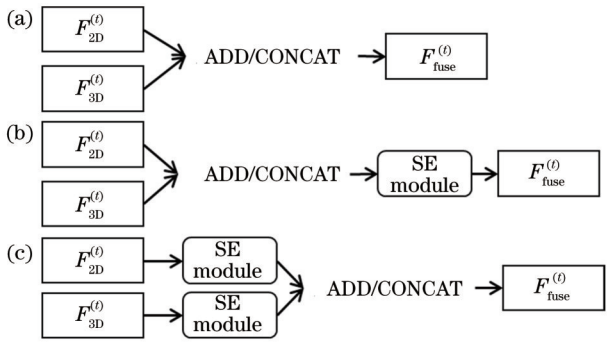

图9 多模态网络中的不同方法。(a)方法C；(b)方法D；(c)方法E

Fig. 9 Different methods in multimodal networks. (a) Method C; (b) method D; (c) method E

不同模态与不同融合方法在KITTI验证集上的各项精度如表1所示。其中，不同模态与不同融合方法在KITTI验证集上的HOTA如图10所示。

通过比较单模态网络和多模态网络的对应方法可知，两种融合方式都提升了网络的性能与指标，这表明两种模态均对最终结果有贡献。

比较特征相加网络与特征拼接网络的对应方法可知，特征拼接在最终指标上有优势。对于HOTA指标，后者在方法C上比前者高0.07个百分点，在方法D上比前者高0.04个百分点，在方法E上比前者高0.12个百分点。这说明特征拼接不会丢失特征信息，能够充分表达各个模态不同通道所包含的信息。需要注意的是，本文为了公平地比较特征相加和特征拼接的优劣，设置2D特征提取网络和3D特征提取网络的输出维度一致，事实上特征拼接不需要不同模态特征维度完全一致。

表1 不同模态与不同融合方法在KITTI验证集上的精度

Table 1 Accuracy of different modalities and different fusion methods on KITTI verification set

Modality	Method	HOTA /%	M_{AssA} /%
Image	A	62.02	57.00
	B	62.21	57.26
Point	A	70.84	74.15
	B	70.97	74.41
Fusion-add	C	71.20	74.83
	D	71.39	75.28
	E	71.54	75.55
Fusion-cat	C	71.27	74.98
	D	71.43	75.36
	E	71.66	75.79

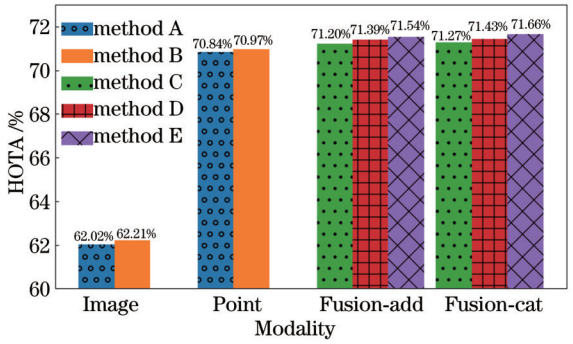

图10 不同模态与不同融合方法在KITTI验证集上的HOTA

Fig. 10 HOTA of different modalities and different fusion methods on KITTI verification set

通过观察图10所示的各个模态中不同方法的指标可以看出：在只使用图像信息的网络中，方法B的HOTA比方法A高0.19个百分点；在只使用点云信息的网络中，方法B的HOTA比方法A高0.13个百分点；在特征相加的网络中，方法D和方法E的HOTA分别比方法C高0.19个百分点和0.34个百分点；在特征拼接的网络中，方法D和方法E的HOTA分别比方法C高0.16个百分点和0.39个百分点。因为注意力机制会强调融合特征中更重要的通道，所以加强了模型的识别能力。

观察特征相加网络和特征拼接网络模型中方法D和方法E的指标可知，方法E的性能优于方法D。在特征相加的网络中，方法E的HOTA比方法D高0.15

个百分点。在特征拼接的网络中，方法E的HOTA比方法D高0.23个百分点。在方法D中，不同模态特征的通道间可能没有相关性，而SE模块将整个融合特征作为输入，此时可能会学习到不合适的通道间关系。然而，方法E在各个模态特征上通过SE模块学习到同一模态特征的通道间关系，并以此来调整各个通道的权重，这增加了特征的表达能力。

3.4.2 相似度估计网络

为了验证相似度估计网络的有效性，选取使用融合方法E的特征拼接网络，分别使用相似度估计网络、欧氏距离和余弦相似度计算相邻帧目标的距离或者相似度。

当使用欧氏距离或者余弦相似度来表达目标特征之间的距离或者相似度时，需要手动设置阈值大小。本组实验中将欧氏距离阈值分别设置为0.5，5.0，50.0，此组阈值是根据经验和多次实验调试得到的。当目标之间的距离小于阈值时，表示两个目标ID一致，而当距离大于阈值时，表示两个目标ID不一致。将余弦相似度阈值分别设置为0.5，0，−0.5，此组阈值是具有代表性的角度余弦值。将相似度估计网络阈值设置为0.5，当目标之间相似度大于此阈值时，表示两个目标ID一致。需要注意的是，此阈值不是调试得来的，它有具体的含义，即表示目标ID一致和ID不一致的概率相等。不同相似度估计方法在KITTI验证集上的精度如表2所示。

通过观察表2可知，不同的阈值使得使用欧氏距离和余弦相似度作为距离度量和相似度度量的组最终的结果不同，并且精度均低于使用相似度估计网络的组。因为通过网络提取的特征通常含有抽象的语义信息，所以相似度估计网络更适合判断该类型特征之间的关系。

表2 不同相似度估计方法在KITTI验证集上的精度

Table 2 Accuracy of different similarity estimation methods on KITTI verification set

Method	Threshold	HOTA /%	M_{AssA} /%
Euclidean distance	0.5	54.84	44.91
	5.0	55.41	46.11
	50.0	54.85	44.63
Cosine similarity	0.5	52.73	41.70
	0.0	54.26	44.15
	−0.5	52.42	41.23
Similarity estimation network	0.5	71.66	75.79

3.5 定性分析

本文选取使用方法E的特征拼接网络作为本文算法，使用方法A的图像单模态网络作为对比算法1，使用方法A的点云单模态网络作为对比算法2，使用方法D的特征拼接网络作为对比算法3，使用方法E的特征相加网络作为对比算法4，mmMOT[8]作为对比算法5，以更直观地分析优劣。其中，mmMOT使用了目标的图片和点云信息，按照同维度相加的方式融合2D特征和3D特征，使用注意力机制调整特征通道权重，并可跨帧关联。图11和图12分别选取了验证集中的部分场景，通过车辆自身视角和鸟瞰图展示了所提算法和对比算法的跟踪结果。

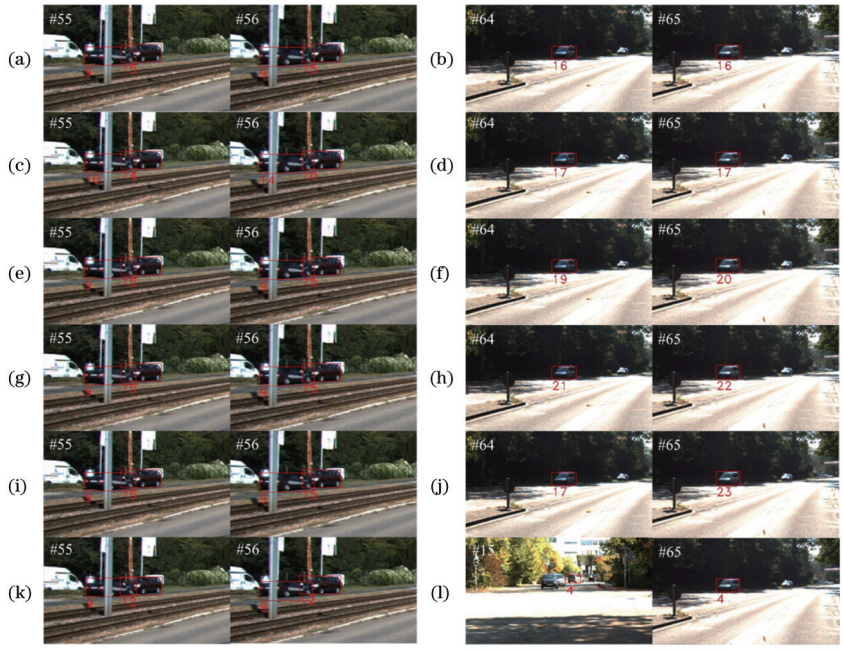

图11 验证集跟踪结果对比。(a)(b)所提算法；(c)(d)对比算法1；(e)(f)对比算法2；(g)(h)对比算法3；(i)(j)对比算法4；(k)(l)对比算法5

Fig. 11 Tracking result comparison on verification set. (a)(b) Proposed algorithm; (c)(d) comparison algorithm 1; (e)(f) comparison algorithm 2; (g)(h) comparison algorithm 3; (i)(j) comparison algorithm 4; (k)(l) comparison algorithm 5

图 11 中每一张图片左上角的数字代表此图片在时间序列中的序号。从图 11(c)中可以看出，由于第 55 帧中 10 号和 14 号目标外观相近，故在第 56 帧中二者被分配了对方的 ID。这是因为对比算法 1 只使用了目标的 2D 特征，当两个目标外观非常相似时，跟踪器不能正确地区分两个目标，进而导致错误的关联。这种情况下，点云表示的 3D 特征会非常重要，所提算法和其他对比算法由于融合了 3D 特征，故不会在两个目标仅外观相似时产生错误关联，如图 11(a)、(e)、(g)、(i)中对应的 9 号和 15 号目标 ID，以及图 11(k)中对应的 9 号和 12 号目标 ID 均被保持了下来，并关联了正确的目标。

图 11 中第 64 帧和第 65 帧中的车辆在远处，点云数量少，3D 特征不充分，此时目标的 2D 特征就非常重要。对比算法 1 只考虑了目标的 2D 特征，正确地跟踪了此目标，如图 11(d)所示。对比算法 2 仅使用了点云特征，进而不能够正确关联，如图 11(f)所示。对比算法 3 和对比算法 4 因为融合了表达不充分的 3D 特征，故不能有选择性地侧重不同的特征，进而导致跟踪失败，如图 11(h)和图 11(j)所示。对比算法 5 虽然在 64 帧(未在图中)和 65 帧中正确关联了目标，但是该 ID 早在第 15 帧时就已被分配出去，即错误地关联了早期目标与当前目标，如图 11(l)所示。所提算法使用注意力机制强调了更为重要的特征，所以这种情况下能够成功地跟踪目标，如图 11(b)所示。

图 12 是两个目标在一段时间序列中所处位置的鸟瞰图，相同形状表示算法认为 ID 一致的目标，方框出现的位置表示跟踪器发生错误跟踪的位置。在图 12(a)中，两个目标在时间序列中被分配为相同的 ID，这是正确的。然而，在图 12(b)中，下方的目标出现了不同的形状，这表明跟踪器为出现过的目标分配了新的 ID，这是跟踪器错误地认为目标不匹配造成的。在图 12(c)～(e)中，上方的目标除了被分配新的 ID 之外，还出现了一个前两种算法没有出现的目标，这表明跟踪器将历史目标的 ID 分配给了新的目标，这是跟踪器错误地匹配目标造成的。在图 12(f)中，下方的轨迹中有一段空白，mmMOT 错误地认为目标不是车辆，左方与上方的符号一样，轨迹却不连贯，这是 mmMOT 错误地关联造成的。所提算法由于融合了 2D 特征和 3D 特征，并且使用注意力机制强调重要的特征，因此在出现各种异常情况时，依然能够正确地区分和识别目标并且分配正确的 ID。

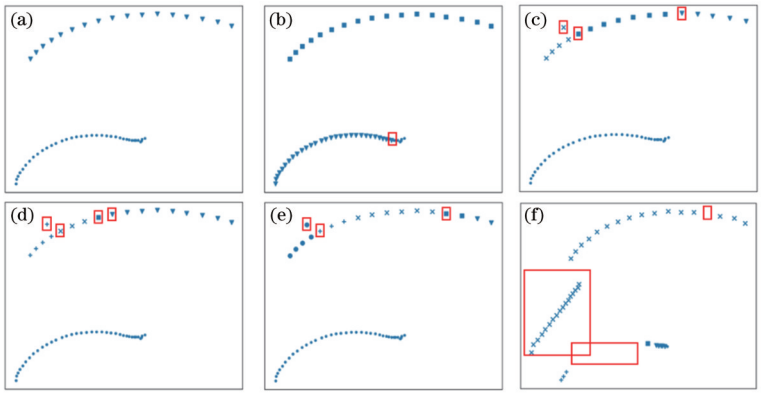

图 12 验证集跟踪结果的鸟瞰图。(a)所提算法；(b)对比算法 1；(c)对比算法 2；(d)对比算法 3；(e)对比算法 4；(f)对比算法 5

Fig. 12 Bird's eye view of tracking result comparison on verification set. (a) Proposed algorithm; (b) comparison algorithm 1; (c) comparison algorithm 2; (d) comparison algorithm 3; (e) comparison algorithm 4; (f) comparison algorithm 5

3.6 测试集结果

所提算法在 KITTI 测试集上的结果如表 3 所示。可以发现，相比最近的一些公开工作，所提算法在 HOTA 上有一些优势。Complexer-YOLO[19]在检测端先对图片进行语义分割，再对对应的点云信息指定不同的类，并且在跟踪端仅使用了目标的 3D 位置信息。CIWT[20]使用立体图片作为输入，同时使用了图片信息和深度信息，跟踪时同时估计和更新目标的 2D 位置和 3D 位置，以解决远处物体在激光雷达中消失的问题。Point3DT[21]仅使用了点云信息。mmMOT 使用了目标的图片和点云信息，按照同维度相加的方式融合 2D 特征和 3D 特征，并使用注意力机制调整特征通道权重。Quasi-Dense[22]和 SRK_ODESA[23]仅使用图片信息作为输入，Quasi-Dense 没有对目标周围的框进行非极大值抑制类似处理，所有框都会被网络学习来表示目标的特征和相似度，而 SRK_ODESA 着重学习目标外观上更具区分度的描述符。MOTFusion[24]使用立体图片作为输入，同时使用了 2D 特征和 3D 特征，并且算法需要在获取全部时刻下各个目标的信息后才进行关联。所提算法在检测精度低于部分算法的情况下，能获得更高的关联精度和 HOTA，说明所提算法对检测质量的依赖更低，能更好地提取目标特征，减少错误关联，最终实现更高的精度。

4 结 论

提出一种基于多模态特征融合与可学习式目标相似度估计的 3D 多目标跟踪算法，分析并验证了不同特征融合方式和不同注意力模块放置位置的区别。在 KITTI 数据集上的实验结果表明，多模态特征融合之后的表达能力优于单一模态特征，并且注意力机制能

表3 不同算法在KITTI测试集上的精度

Table 3 Accuracies of different algorithms on KITTI test set

Algorithm	HOTA /%	M_{AssA} /%	M_{DetA} /%
Complexer-YOLO	49.12	39.34	62.44
CIWT	54.90	49.99	60.57
Point3DT	57.20	59.15	55.71
mmMOT	62.05	54.02	72.29
Quasi-Dense	68.45	65.49	72.44
SRK_ODESA	68.51	65.49	75.40
MOTFusion	68.74	66.16	72.19
Proposed	69.24	68.46	70.71

够调整特征通道的权重使其表达能力更强。所提的相似度估计网络相比于欧氏距离能够更好地利用高维度信息,分类更准确,并且无需手动设置相似度阈值。所提算法对检测结果完全信任,并且只考虑了相邻帧的目标关联。下一步将设计合适的模块过滤误检,并实现跨帧关联。

参 考 文 献

[1] 陈志旺, 张忠新, 宋娟, 等. 基于目标感知特征筛选的孪生网络跟踪算法[J]. 光学学报, 2020, 40(9): 0915003.
Chen Z W, Zhang Z X, Song J, et al. Tracking algorithm for Siamese network based on target-aware feature selection[J]. Acta Optica Sinica, 2020, 40(9): 0915003.

[2] 李畅, 杨德东, 宋鹏, 等. 基于全局感知孪生网络的红外目标跟踪[J]. 光学学报, 2021, 41(6): 0615002.
Li C, Yang D D, Song P, et al. Global-aware Siamese network for thermal infrared object tracking[J]. Acta Optica Sinica, 2021, 41(6): 0615002.

[3] 刘宗达, 董立泉, 赵跃进, 等. 视频中快速运动目标的自适应模型跟踪算法[J]. 光学学报, 2021, 41(18): 1815001.
Liu Z D, Dong L Q, Zhao Y J, et al. Adaptive model tracking algorithm for fast-moving targets in video[J]. Acta Optica Sinica, 2021, 41(18): 1815001.

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

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

[6] Chiu H K, Prioletti A, Li J, et al. Probabilistic 3D multi-object tracking for autonomous driving[EB/OL]. (2020-01-16)[2020-10-09]. https://arxiv.org/abs/2001.05673.

[7] Weng X S, Wang J R, Held D, et al. 3D multi-object tracking: a baseline and new evaluation metrics[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, October 24-January 24, 2021, Las Vegas, NV, USA. New York: IEEE Press, 2021: 10359-10366.

[8] Zhang W W, Zhou H, Sun S Y, et al. Robust multi-modality multi-object tracking[C]// IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 2365-2374.

[9] Vora S, Lang A H, Helou B, et al. PointPainting: sequential fusion for 3D object detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 4603-4611.

[10] Shenoi A, Patel M, Gwak J Y, et al. JRMOT: a real-time 3D multi-object tracker and a new large-scale dataset [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, October 24-January 24, 2021, Las Vegas, NV, USA. New York: IEEE Press, 2021: 10335-10342.

[11] Huang K, Hao Q. Joint multi-object detection and tracking with camera-LiDAR fusion for autonomous driving[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems, September 27-October 1, 2021, Prague, Czech Republic. New York: IEEE Press, 2021: 6983-6989.

[12] Chiu H K, Li J, Ambruș R, et al. Probabilistic 3D multi-modal, multi-object tracking for autonomous driving[C]//IEEE International Conference on Robotics and Automation, May 30-June 5, 2021, Xi′an, China. New York: IEEE Press, 2021: 14227-14233.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10) [2021-10-15]. https://arxiv.org/abs/1409.1556.

[14] Charles R Q, Hao S, Mo K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 77-85.

[15] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[16] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]//IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 3354-3361.

[17] Luiten J, Os Ep A A, Dendorfer P, et al. HOTA: a higher order metric for evaluating multi-object tracking [J]. International Journal of Computer Vision, 2021, 129 (2): 548-578.

[18] Bhattacharyya P, Huang C J, Czarnecki K. SA-Det3D: self-attention based context-aware 3D object detection [EB/OL]. (2021-08-19)[2021-10-15]. https://arxiv.org/abs/2101.02672.

[19] Simon M, Amende K, Kraus A, et al. Complexer-YOLO: real-time 3D object detection and tracking on semantic point clouds[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, June 16-17, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 1190-1199.

[20] Osep A, Mehner W, Mathias M, et al. Combined image- and world-space tracking in traffic scenes[C]// IEEE International Conference on Robotics and Automation, May 29-June 3, 2017, Singapore. New York: IEEE Press, 2017: 1988-1995.

[21] Wang S K, Sun Y X, Liu C J, et al. PointTrackNet: an end-to-end network for 3-D object detection and tracking from point clouds[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3206-3212.

[22] Pang J M, Qiu L L, Li X, et al. Quasi-dense similarity learning for multiple object tracking[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 164-173.

[23] Mykheievskyi D, Borysenko D, Porokhonskyy V. Learning local feature descriptors for multiple object tracking[M]//Ishikawa H, Liu C L, Pajdla T, et al. Computer vision-ACCV 2020. Lecture notes in computer science. Cham: Springer, 2021, 12623: 558-575.

[24] Luiten J, Fischer T, Leibe B. Track to reconstruct and reconstruct to track[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 1803-1810.