计算机科学
COMPUTER SCIENCE

基于图交互与场景感知融合的轨迹预测方法

方阳，赵婷，刘期烈，贺侗，孙开伟，陈前斌

引用本文

方阳，赵婷，刘期烈，贺侗，孙开伟，陈前斌. 基于图交互与场景感知融合的轨迹预测方法[J]. 计算机科学, 2022, 49(10): 258-264.

FANG Yang, ZHAO Ting, LIU Qi-lie, HE Dong, SUN Kai-wei, CHEN Qian-bin. Trajectory Prediction Method Based on Fusion of Graph Interaction and Scene Perception[J]. Computer Science, 2022, 49(10): 258-264.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000172

基于图交互与场景感知融合的轨迹预测方法

方 阳[1] 赵 婷[2] 刘期烈[2] 贺 侗[3] 孙开伟[1] 陈前斌[2]
1 重庆邮电大学计算机科学与技术学院 重庆 400065
2 重庆邮电大学通信与信息工程学院 重庆 400065
3 韩国科学技术院(KAIST)电气工程学院 大田 34141
(fangyang@cqupt.edu.cn)

摘 要 在自动驾驶中,精确的环境感知和对周围交通参与者的轨迹预测对道路安全至关重要。基于此,提出了基于鸟瞰图(Bird Eye View,BEV)的实时端到端轨迹预测框架来同时学习交互和场景信息。该框架主要由图交互网络和金字塔感知网络两个模块组成,前者通过时空图卷积网络对交通参与者之间的交互模式进行编码,后者采用时空金字塔网络对周围信息进行场景建模以获取场景特征。然后,对交互特征和场景特征进行单一尺度融合,从而进行分类和轨迹预测任务。在大规模开源数据集 NuScenes 上的实验和分析表明,与当前先进算法(MotionNet)相比,所提框架平均类别准确度提高了 3.1%,轨迹预测平均误差在行驶速度>5m/s 时降低了 1.43%。此实验结果表明,所提模型具有更好的泛化性和鲁棒性,更符合实际自动驾驶环境中的轨迹预测需求。

关键词:轨迹预测;时空图卷积;时空金字塔;图交互编码;特征融合
中图法分类号 TP183

Trajectory Prediction Method Based on Fusion of Graph Interaction and Scene Perception

FANG Yang[1],ZHAO Ting[2],LIU Qi-lie[2],HE Dong[3],SUN Kai-wei[1] and CHEN Qian-bin[2]
1 School of Computer Science and Technology,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
2 School of Communication and Information Engineering,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
3 School of Electrical Engineering,KAIST,Daejeon 34141,South Korea

Abstract To accurately perceive the environment and predict the trajectory of the surrounding traffic participants for autonomous driving,we propose a real-time end-to-end trajectory prediction framework based on bird eye view(BEV) to learn both interaction and scene information simultaneously. The framework consists of two essential modules:graph interaction network and pyramid perception network. The former encodes the interaction patterns among traffic participants through a spatiotemporal graph convolutional network,and the latter adopts a spatiotemporal pyramid network to model the surrounding information and obtain the scene features. Next,interactive features and scene features are fused at a unified scale to perform classification and trajectory prediction tasks. Experiments and analysis on Nuscenes,a large open-source dataset,indicate that the proposed framework achieves a higher classification accuracy of 3.1% and 1.43% less predicted trajectory loss than MotionNet. Hence,our framework outperforms state-of-the-art algorithms in terms of generalization and robustness,and is more in line with perception requirements in actual autonomous driving scenes.

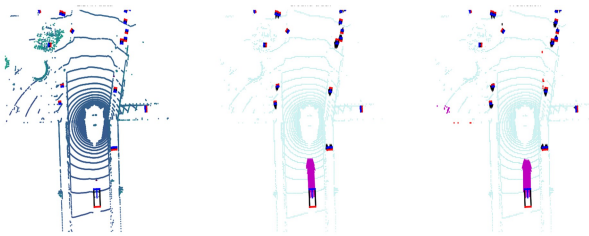
Keywords Trajectory prediction,Spatiotemporal graph convolutional,Spatiotemporal pyramid,Graph interaction encoding,Feature fusion

到稿日期:2021-10-25 返修日期:2022-02-28
基金项目:重庆市教委青年项目(KJQN202100634);重庆市科技创新领军人才支持计划(CSTCCXLJRC201908);重庆市自然基金重点项目(cstc2019jcyj-zdxm0008);重庆市教委重点项目(KJZD-K201900605);国家自然科学基金青年科学基金项目(61806033);重庆市自然科学基金面上项目(cstc2019jcyj-msxmX0021);"成渝地区双城经济圈建设"科技创新项目(KJCXZD2020027)
This work was supported by the Science and Technology Research Program of Chongqing Municipal Education Commission(KJQN202100634),Chongqing Science and Technology Innovation Leading Talent Support Program(CSTCCXLJRC201908),Basic and Advanced Research Projects of CSTC(cstc2019jcyj-zdxmX0008),Science and Technology Research Program of Chongqing Municipal Education Commission(KJZD-K201900605),Young Scientists Fund of the National Natural Science Foundation of China(61806033),Natural Science Foundation of Chongqing,China(cstc2019jcyj-msxmX0021) and Scientific and Technological Innovation Projects of the Construction of the Two Cities Economic Circle in Chengdu Chongqing Region(KJCXZD2020027).
通信作者:赵婷(S190101071@stu.cqupt.edu.cn)

1　引言

近年来，人工智能在自动驾驶领域的应用取得了前所未有的进展，越来越多的智能算法也逐渐融入我们日常驾驶中，成为不可或缺的一部分。减少交通事故，提高交通道路安全，是自动驾驶车辆的目的之一[1]。为了使自动驾驶车辆在路上安全有效地行驶，通过过去周围环境的状态来预测接下来附近交通参与者的未来轨迹至关重要。最近，相关研究成为了学术界和工业界的研究重点。自动驾驶车辆需要通过周围交通参与者的当前状态和其未来的行为轨迹来正确规划路线[2-3]，所以，感知环境的研究主要包括两个方面：1）感知，即从背景中识别出前景目标；2）轨迹预测，即预测目标未来的行为轨迹。因为自动驾驶车辆没有传统的驾驶人员，所以在极短时间内对周围车辆或行人的运动状态和未来轨迹进行预知是非常困难的。同时，交通参与者在未来某一段时间的行驶轨迹不仅依赖于当前自身的行驶状态，还与其他交通参与者之间的相互影响有关[4]。早期的轨迹预测研究大部分只考虑了周围环境状态，并未考虑交通参与者相互之间的影响。基于此观察，本文提出了相关方案来描述自动驾驶车辆周围的交通参与者，如车辆之间存在的特定的交互关系。

早期感知环境的方法大部分依赖于目标检测器，但是目标检测器无法识别训练集中从未出现过的物体，从而无法预测相应的未来轨迹。基于上述问题，本研究通过采用鸟瞰图[5]和 Occupancy Grid Map（OGM）的方式来表示周围环境状态，根据驾驶场景中周围交通参与者的位置关系特性，采用图卷积对周围交通参与者的特性进行建模，其中图节点为交通参与者，边表示交通参与者之间的交互。同时，通过图交互网络和金字塔感知网络的融合来对自动驾驶车辆周围交通参与者的未来轨迹进行更准确的预测，更好地为自动驾驶车辆进行轨迹规划。图 1 给出了本文方法的可视化结果。其中，图 1(a)是激光雷达点云图，不同的箭头颜色表示不同的交通参与者类别，箭头方向和长度表示运动方向和距离，中间的白色车状物体表示自动驾驶车辆，周围的框代表其周围的交通参与者；图 1(b)表示在鸟瞰图下，场景中交通参与者的类别、运行方向以及轨迹真值；图 1(c)表示预测的交通参与者的类别、运动方向及轨迹。



(a)激光雷达点云　(b) 轨迹真实值　(c) 轨迹预测值

图 1　目标分类与轨迹预测的可视化结果（电子版为彩图）

Fig. 1　Visualized result of object detection and trajectory prediction

本研究的主要贡献为以下 3 个方面：

（1）提出了用时空图卷积来表示自动驾驶车辆周围交通参与者的交互关系，通过图表达的交通参与者的交互特征来准确预测周围交通参与者的轨迹，辅助自动驾驶车辆进行准确的规划路径。

（2）提出了一种融合图神经网络生成的交互特征和金字塔感知网络生成的场景特征的框架，其能够有效获取空间、时间和交互关系 3 方面的信息。

（3）通过在大规模实际驾驶场景数据集 NuScenes 上进行实验和分析，验证了所提出的融合交互特征和场景特征的框架相比当前先进算法具有一定的预测精度优势。

2　相关工作

2.1　环境感知

近年来，研究者们提出了独立或联合的方法来解决环境感知和动作预测问题[6-8]。传统的环境感知方法是基于相机图片的二维目标检测[9]，或者是基于激光雷达点云[10]的三维目标检测，又或者是基于二者融合的检测[11]，其中基于激光雷达点云的轨迹预测方法取得了显著进展。Lefevre 等[12]回顾了各种基于物理模型和传统机器学习算法的传统方法，如隐马尔可夫模型、支持向量机和动态贝叶斯网络，这些方法主要依赖于边界框检测。文献[11,13]将两阶段检测器[14]应用于端到端的框架中，直接生成边界框和未来轨迹。两阶段检测器包含利用区域建议网络（Region Proposal Network，RPN）来学习潜在目标所在的感兴趣区域（Region Of Interest，ROI），并根据目标检测获得边界框，然后将边界框输入到目标跟踪器中来进行动作估计和轨迹预测。上述方法跳过了一阶段检测器中 Region Proposals 的生成步骤，直接学习一个生成物体边界框的网络。这些方法都依赖于目标检测器，但是目标检测器无法识别训练集中从未出现过的物体，从而导致无法预测其未来轨迹。

Schreiber 等[15]提出了 OGM 方法来解决上述问题。OGM 将自动驾驶车辆周围的空间划分为相同的单元来代表周围区域的占用状态，即占用或者空闲。但是，OGM 没有纳入时间维度，很难建模对象的非线性动态行为；而且 OGM 没有考虑分类任务。文献[16]通过计算鸟瞰图中的多通道特征图和正面视图中的柱体坐标，引入了激光雷达点云的多视图表示来进行分类和目标检测。文献[17]使用特征编码器将点云投影到稀疏的 BEV 中，进行体素化，通过二维卷积提取特征。Cui 等[18]使用深度学习方法创建 BEV 光栅，通过编码高清地图和环境来预测周围物体的未来轨迹。Xu 等[19]将原始点云转换为 BEV 图，然后将 BEV 图和二进制图片融合输入到卷积神经网络中以提取深层特征，从而进行轨迹预测。

此外，基于类时空金字塔感知网络的自动驾驶环境感知方法发展迅速。Zeng 等[20]提出时空卷积模型将 LiDAR 点云和高清地图作为输入，以 3D 检测的形式生成可解释的中间表示，并在规划范围内预测交通参与者的可能运行轨迹。Casas 等[21]构建了一个多时空 3D 网络，该网络同时对多维 Lidar 点云图和车道、交叉路口、交通灯等具有先验知识的语义元素进行建模，实现自动驾驶的检测和轨迹预测功能。文献[22]提出了一种新颖的端到端两阶段网络，即时空交互网络，实现对交通参与者的 3D 几何信息和时间信息的建模，从而更好地完成下游的分类和检测任务。然而，这些轨迹预测方法都没有充分考虑交通参与者之间直接或者间接的交互影响。

2.2 图卷积模型

图(Graph)是一种数据结构,图中的节点表示网络中的个体,边表示个体之间的连接关系。三维卷积神经网络(3-Dimension Convolutional Neural Network, 3D-CNN)[23]虽然同样可以用于位置建模,但是其计算成本高,不能处理任意图结构。图卷积网络(Graph Convolutional Networks, GCNs)[24]可以对任意图中不同节点之间的依赖关系进行建模并传播消息,因而受到越来越多的关注,并被成功地应用于各种计算机视觉任务中[25-26]。本文将图卷积网络引入自动驾驶场景来完成轨迹预测任务。

3 系统模型与问题描述

3.1 场景特征

准确感知周围环境对轨迹预测至关重要,然而大部分环境感知的方法是基于特征边界的提取和检测,对于复杂场景鲁棒性不高,很难用于实际的自动驾驶场景。同时,因为原始的激光雷达数据是稀疏且不规则的三维数据,不能直接进行标准的二维卷积运算,通过三维卷积来提取特征会耗费大量的内存并且导致计算量庞大和耗时过长,所以相关工作采用了点云投影到二维的方法来描述场景。另外,激光雷达的位移,导致每帧点云的原始坐标都是对应时刻的局部坐标,因此我们需要通过坐标变换来将已有点云统一到一个世界坐标系下。本工作采用了文献[27]提出的 BEV 表示法,将 3D 点云转换为 2D 伪图像,将三维的高度信息映射为图像的通道。具体方法为将 3D 点云离散为规则的体素格,其中将有点云的体素标记为 1,没有点云的体素标记为 0。本方法对每一个网格进行分类,并且根据当前和之前的状态来预测接下来的轨迹。获取局部和全局上下文信息在预测任务中至关重要,即何时聚合时间特征,何时提取多尺度时空特征。如图 2 所示,本文采用了修改的时空金字塔网络[26]作为提取场景特征的基础网络。时空金字塔网络主要是使用卷积块来进行特征提取,1 个卷积块包含 1 个二维卷积和 1 个核函数为$(k,1,1)$的三维卷积。在空间维度上,缩放步长为 2 来获取多尺度空间特征;在时间维度上,使用一维卷积来提取时间特征,并且通过降低时间分辨率来提取多尺度时间特征。将经过时间池化的高层语义时间特征与相对应层的特征进行融合,得到不同尺度的时空特征。

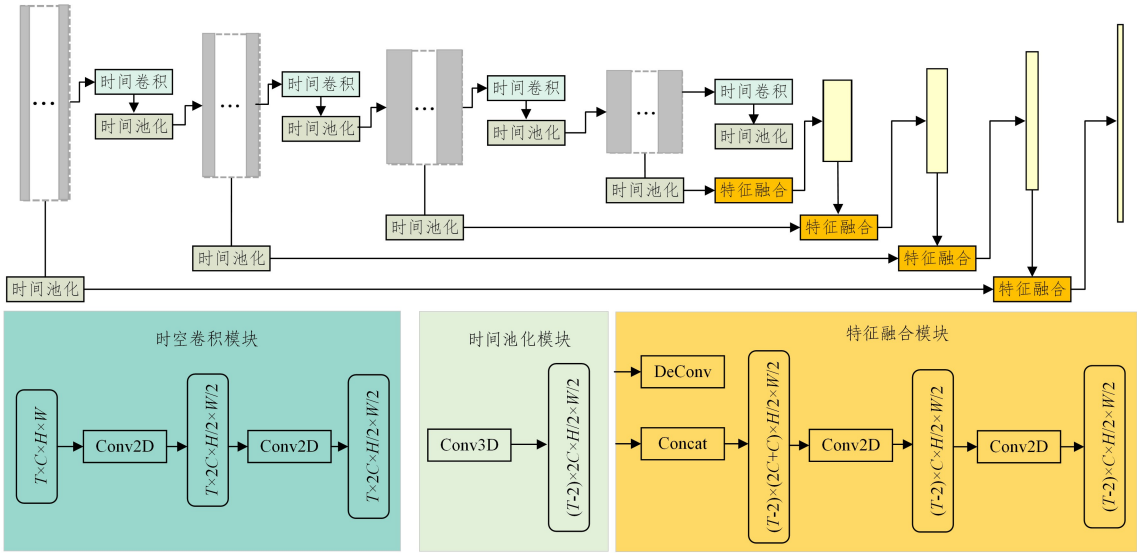


图 2 时空金字塔网络

Fig. 2 Spatiotemporal pyramid network

3.2 交互特征

交通参与者行为之间存在相互依赖性,因此预测自动驾驶车辆周围交通参与者的轨迹需要参考其他交通参与者的信息。目前,大部分已有研究没有考虑周围交通参与者之间的相互影响。基于此,考虑到图卷积可以建模社交关系,我们采用了图卷积来建模周围交通参与者之间的相互影响。首先对数据集中的数据进行预处理,得到每一个场景下每一帧的周围交通参与者的坐标信息;随后用周围交通参与者的轨迹信息构造一组表示场景中周围交通参与者在每个时间步长 t 的相对位置的空间图 G_t。如图 3(a)所示,G_t 的定义为 $G_t=(V_t,E_t)$,其中 $V_t=\{v_t^i \mid \forall i\in\{1,\cdots,N\}\}$ 代表了图 G_t 的节点,位置信息(x_t^i,y_t^i)即是 V_t^i 的值。E_t 是图 G_t 的边,表示为 $E_t=\{e_t^{ij} \mid \forall i,j\in\{1,\cdots,N\}\}$。$v_t^i$ 和 v_t^j 如果相连,则 $e_t^{ij}=1$;如果不相连,则 $e_t^{ij}=0$。同时,为了建模两个节点之间的相互影响程度,我们使用了 a_t^{ij},其由每个 e_t^{ij} 的核函数计算得到,a_t^{ij} 被加在加权邻接矩阵 $\boldsymbol{A}_t$ 中。最后,定义 $a_{\mathrm{sim},t}^{ij}$ 作为核函数:

$$a_{\mathrm{sim},t}^{ij}=\begin{cases}1/\parallel v_t^i-v_t^j\parallel_2, & \parallel v_t^i-v_t^j\parallel_2\neq 0\\ 0, & \parallel v_t^i-v_t^j\parallel_2=0\end{cases} \tag{1}$$

图卷积运算公式如下:

$$v^{i(l+1)}=\sigma\left(\frac{1}{\Omega}\sum_{v^{j(l)}\in B(v^{i(l)})}p(v^{i(l)},v^{j(l)})\cdot w(v^{i(l)},v^{j(l)})\right) \tag{2}$$

其中,σ 是激活函数;$\frac{1}{\Omega}$ 是归一化函数;$B(v^i)=\{v^j \mid d(v^i,v^j)\leqslant D\}$ 是顶点 v^i 的邻居集,$d(v^i,v^j)$表示连接 v^i 和 v^j 的最短路径,D 表示路径集;p 是抽样函数;w 是权重因子。

将图表示的周围交通参与者关系信息输入图卷积,考虑其时间维度。我们定义一个新图 G,如图 3(b)所示。新图 G 是 G_t 的属性集合,G 包含周围交通参与者轨迹的时空信息,$G_1,\cdots,G_T$ 的拓扑结构完全一致。$G=(V,E)$,其中 $V=\{v^i \mid i\in\{1,\cdots,N\}\}$,$E=\{e^{ij} \mid \forall i,j\in\{1,\cdots,N\}\}$。图 G 中的 v^i 是

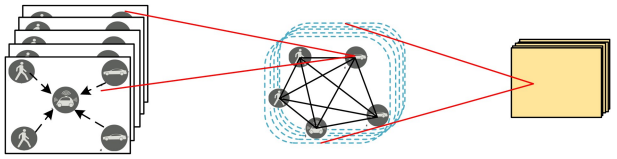
v_t^i 的集合，$\forall t \in \{0,\cdots,T\}$。同时，$G$ 的加权邻接矩阵 $\boldsymbol{A}$ 是 $\{\boldsymbol{A}_1,\cdots,\boldsymbol{A}_T\}$ 的集合。要使模型正确实施，我们需要归一化邻接矩阵，使用下面的公式均匀地归一化每个 $\boldsymbol{A}_t$。

$$\boldsymbol{A}_t=\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\hat{\boldsymbol{A}}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}} \tag{3}$$

其中，$\hat{\boldsymbol{A}}_t=\boldsymbol{A}_t+\boldsymbol{I}$，$\boldsymbol{\Lambda}_t$ 是对角节点度矩阵，$\hat{A}$ 和 Λ 分别表示 $\hat{\boldsymbol{A}}_t$ 和 $\boldsymbol{\Lambda}_t$ 的集合。处在时间 t 和网络层 l 的节点记为 $V_t^{(l)}$。$V^{(l)}$ 是 $V_t^{(l)}$ 的集合。根据上面的图卷积计算公式，可以把时空图卷积网络表示为：

$$f(V^{(l)},\boldsymbol{A})=\sigma(\Lambda^{-\frac{1}{2}}\hat{A}\Lambda^{-\frac{1}{2}}V^{(l)}\boldsymbol{W}^{(l)}) \tag{4}$$

经过时空图卷积网络之后，可获得交通参与者的交互信息特征，如图 3(c)所示。



(a)空间图 G_t　　(b) 关系新图 G　　(c) 交互信息特征图

图 3　GCNN 模型结构示例

Fig. 3　GCNN model structure

3.3　模型融合

图 4 简要描述了图卷积模型与时空金字塔模型的融合框架。首先，通过图卷积模型得到交通参与者之间的交互特征，并且通过时空金字塔模型获取周围信息的场景特征。然后，将交互特征和场景特征进行单一尺度融合，并通过一个反卷积层将融合特征缩放至原始输入尺寸，得到分类和轨迹预测的结果。具体操作步骤如下：首先对周围物体进行分类，将 BEV 图分割后对每一个单元格分类，用两层二维卷积实现，分类结果的输出维度为 $H\times W\times C$；然后进行状态估计，估计每个单元格的运动状态(静态或者动态)，为轨迹预测提供辅助信息。状态估计可以抑制轨迹预测的抖动，因为即使是背景物体，也可能有微小运动，状态估计可以设置阈值来使背景物体为静态，其输出维度是 $H\times W$；最后是轨迹预测，预测未来的单元格的位置，预测的单元表示为 $\{P^\tau\}_{\tau=t}^{t+N}$，其中 $P^\tau\in\mathbb{R}^{X\times W\times 2}$ 表示物体在 τ 时刻的位置，t 是当前时刻，N 是未来帧数，输出维度为 $N\times H\times W\times 2$。

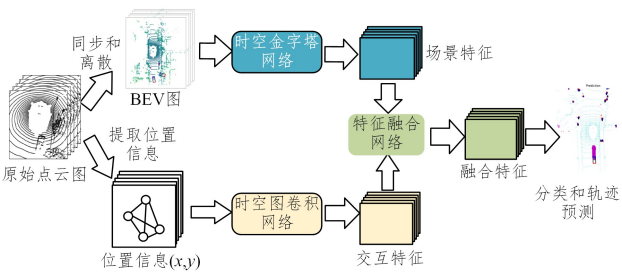


图 4　特征融合系统框架图

Fig. 4　Frame diagram of feature fusion system

3.4　损失函数

本文采用带权重的交叉熵损失函数来处理分类输出，并且通过不同的权值来解决每个类别的类别不平衡问题。基于文献[27]的分析，对于轨迹预测，采用加权 smooth L1 损失函数，其中权重的设定遵循分类的权重设置。然而，上述损失只能保证训练的全局规范化，不能保证局部的时空一致性，因此采用了时空一致性损失。空间一致性损失函数表示为：

$$L_s=\sum_k\sum_{(i,j),(i',j')\in o_k}\|P_{i,j}^{(t)}-P_{i',j'}^{(t)}\| \tag{5}$$

其中，$\|\cdot\|$ 表示 smooth L1 损失，O_k 表示索引为 k 的对象实例，$P_{i,j}^{(t)}\in R^2$ 表示在时间 t 和位置(i,j)时的预测信息，$P_{i',j'}^{(t)}\in R^2$ 表示在时间 t 和位置(i',j')时的真值信息。考虑到连续坐标系之间不会有剧烈的运动变化，时间一致性损失函数与空间一致性损失大体一致，因此可以通过以下公式来实现局部时间上的时间约束：

$$L_t=\sum_k\|P_{o_k}^{(t)}-P_{o_k}^{(t+\Delta t)}\| \tag{6}$$

其中，$P_{O_k}^{(t)}$ 表示目标 k 的总体运动，本文用平均运动来表示，即 $P_{o_k}^{(t)}=\sum_{(i,j)\in o_k}P_{i,j}^{(t)}/M$，$M$ 是单元格的数量。

4　实验与结果分析

4.1　数据集和评价标准

NuScenes 是提供自动驾驶车辆全套传感器数据的大型数据集，包括 6 个相机、1 个激光雷达、5 个毫米波雷达以及 GPS 和 IMU。数据集有 1 000 个带有注释样本的场景，每个场景的采样时长为 20 s，包含各种驾驶情景。本文使用数据集中的激光雷达点云数据。NuScenes 中的标签为目标检测的边界框，并没有提供运动信息，因此我们将两帧之间的运动设定为在边界框内的点云，根据其坐标信息以及相对边界框中心的偏转位移来计算；对于边界框外的点，将其运动置为 0。我们同时对点云进行裁剪，在 x 坐标轴上将范围设定为正、负轴都是 32 m；在 y 坐标轴上也设定了同样的范围；在 z 坐标轴上，考虑到激光雷达传感器放在车辆顶部，将负轴方向设置为 3 m，正轴为 2m。划分每个体素格为长 0.25 m，宽 0.25 m，高 0.4 m。对于轨迹预测，本文计算相邻时间戳的相对位移，同时把所有的单元格根据不同的速度分为 3 组：静态、速度$<$5 m/s，以及速度$>$5 m/s。在每一组中，计算估计位移和地面真值位移之间的平均 L2 范式距离，即平均位移误差(Average Displacement Error，ADE)，如式(7)所示：

$$ADE=\frac{\sum_{n\in N}\sum_t\|\hat{p}_t^n-p_t^n\|_2}{N} \tag{7}$$

其中，N 表示交通参与者的个数，$\hat{p}_t^n$ 表示在 t 时刻第 n 个交通参与者的预测轨迹，p_t^n 表示真值轨迹。

式(8)表示所有单元格的平均分类精度(Overall cell classification Accuracy，OA)：

$$OA=CCC/AC \tag{8}$$

其中，CCC(Correct Classified Cells)表示正确分类的单元格数，AC(All Cells)表示总单元格数。

式(9)表示平均类别准确度(Mean Category Accuracy，MCA)：

$$MCA=[CA(\text{Bg})+CA(\text{Vehicle})+CA(\text{Ped})+CA(\text{Bike})+CA(\text{Others})]/5 \tag{9}$$

其中，CA(Bg)表示背景的分类精度，CA(Vehicle)表示车辆的分类精度，CA(Ped)表示行人的分类精度，CA(Bike)表示自行车的分类精度，CA(Other)表示其他交通参与者的分类精度。

4.2 仿真结果分析

为了验证本文融合方法的性能，将其与其他几种方法进行对比。表1列出了预测未来几帧运动时不同方法的误差。本文整个算法过程耗费时间为0.019 s，比其他对比方法更快，与基线方法MotionNet[27]速度相当，其中图卷积运行时间为4.669×10^{-5} s，增加的时间开销可忽略不计，在周围交通参与者速度<5 m/s和速度>5 m/s的单元格运动速度下比其他方法性能更好。

表1 预测误差的对比

Table 1 Comparison of prediction error

method	static		velocity≤5 m/s		velocity>5 m/s		Infer speed/s
	ADE/m	Median/m	ADE/m	Median/m	ADE/m	Median/m	
Static Model	0	0	0.6111	0.0971	8.6517	8.1412	—
FlowNet3D[28]	0.0410	0	0.8183	0.1783	8.5261	8.0230	0.4340
HPLFlowNet[29]	0.0041	0.0002	0.4458	0.0960	4.4206	2.4881	0.3520
PointRCNN[11]	0.0204	0	0.5514	0.1627	3.9888	1.6252	0.2010
LSTM-EncoderDecoder[31]	0.0358	0	0.3551	0.1044	1.5885	1.0003	0.0420
MotionNet[27]	0.0256	0	0.2565	0.0962	1.0744	0.7332	0.0270
MotionNet[27]+spatiotemporal consistency loss	0.0239	0	0.2467	0.0961	1.0109	**0.6994**	0.0190
Ours	0.0345	0	0.2720	0.0971	1.1074	0.7844	0.0190
Ours +spatiotemporal consistency loss	0.0286	0	**0.2462**	**0.0958**	**0.9966**	0.7177	0.0190

从表1可以看出，在静态场景下，静态模型误差为0，效果最佳，但是静态模型是在极限条件下，并不适用大部分现实场景。FlowNet3D[28]和HPLFlowNet[29]两者在动态环境下的误差比静态模型更大，效果不佳。PointRCNN[11]由于依赖目标检测框，结果也不理想，轨迹预测误差较大。在速度<5m/s的交通场景中，基于目标检测框架中效果较好的LSTM-EncoderDecoder[30]方法的ADE为0.3551，本文所提算法为0.2467，后者比前者低了约11%；在速度>5m/s情况下，本文方法相比LSTM-EncoderDecoder，误差降低了60%。相比之下，本文方法能更准确有效地预测运动，表明其能为运动规划提供辅助信息。与基于边界框的方法相比，本文方法可以更好地感知训练集中未出现的物体。主要原因是基于边界框的方法利用ROI全局信息获取目标，然而这个信息在不同类别对象中存在差异，很难从可见对象泛化到不可见对象。相比之下，本文方法是在网格单元中提取对象类别共享的局部信息，直接对每个网格单元做类别和动作预测，避免了ROI存在的问题，使得预测效果更佳。

同样，使用激光雷达点云的MotionNet是本文获得场景特征所参考的基础，采用其原始模型和加入了时空一致性损失的模型来进行对比。因为本方法未使用多梯度下降算法，所以没有将MotionNet加入了多梯度下降算法的预测结果与本方法进行对比。MotionNet提出了空间一致性损失、前景时间一致性损失和背景时间一致性损失，根据其结果，空间一致性损失和前景时间一致性损失有利于减小误差，因此本文采用了时空一致性损失。在速度<5 m/s情况下，本文所提方法和MotionNet相比误差降低了0.02%。考虑到交通参与者的行驶速度<5m/s通常发生在交通拥堵等非正常行驶环境中，或交通参与者属于行人等少数类别，因此本文主要考虑自动驾驶的一般行驶场景，且交通参与者主要以车辆为主，以速度>5 m/s时的实验性能作为所提模型的主要评测标准。在速度>5 m/s的行驶场景中，基线方法MotionNet的平均误差为1.0109，本文所提方法为0.9966，平均误差降低了1.43%。通过分析得知，误差降低的原因是本文融入了周围交通参与者的交互信息，周围交通参与者的未来轨迹会受到相互之间的影响。与基线MotionNet结果进行对比得知，本文方法效果更佳。表1同样证明了时空一致性损失的有效性。在速度<5 m/s的情况下，本文平均位移误差在未加时空一致性损失时为0.2720，加上时空损失之后为0.2462，误差降低了2.58%。在静态和速度>5m/s的场景下，本文方法的性能也较其他方法有所提高。对于分类任务，本文所提方法相较于当前最先进算法也具有优势。对于自动驾驶车辆周围的交通参与者，车辆和行人占大多数。如表2所列，基线模型MotionNet的车辆分类精确度是90.7%；本文所提方法的分类精确度为91.3%，相比基线模型提高了0.6%。对于行人，本文未加时空一致性的精确度为83.5%，相较于MotionNet提高了7.3%。对于平均类别精度，MotionNet为71.3%，而本文达到了74.4%，相较于MotionNet提升了3.1%，提升效果显著。

表2 分类精度的对比

Table 2 Comparison of classification accuracy

method	Classification accuracy/%						
	background	car	pedestrian	bike	others	MCA	OA
PointRCNN[11]	94.4	78.7	44.1	11.9	44.0	55.4	94.0
LSTM-EncoderDecoder[31]	93.8	91.0	73.4	17.9	71.7	69.6	92.8
MotionNet[27]	97.3	91.1	76.2	20.6	66.1	70.3	96.1
MotionNet[27]+spatiotemporal consistency loss	**97.6**	90.7	77.2	25.8	65.1	71.3	**96.3**
Ours	94.1	91.1	**83.5**	27.9	**75.8**	**74.4**	93.1
Ours + spatiotemporal consistency loss	95.8	**91.3**	80.5	**29.0**	73.6	74.2	94.8

有代表性的4个场景的结果的可视化如图5所示，其中不同颜色表示不同的交通参与者，箭头表示运动方向。蓝色表示背景，紫色表示车辆，黑色表示行人，绿色表示自行车，红色表示其他。图5(a)为直线道路上对左右车辆进行预测，

场景较为简单，通过对比预测的轨迹和真值可以发现本文方法的预测效果较好。图 5(b)中，自动驾驶车辆周围有各种类型的交通参与者，相互之间交互性较强，本文方法由于加入了时空图卷积建模周围交通参与者的交互，因此可以发现每种类型的交通参与者的轨迹预测性能都有所提高。图 5(c)是一个交叉路口，可以看出，对于自动驾驶车辆迎面而来的车辆和路口左右侧车辆，本文方法都能较为准确地预测其轨迹。图 5(d)中，对于其他类别的交通参与者，通过对比预测和真值发现，本文方法仍然能准确预测目标轨迹。综上所述，本文方法能准确分类并输出较为准确的轨迹。

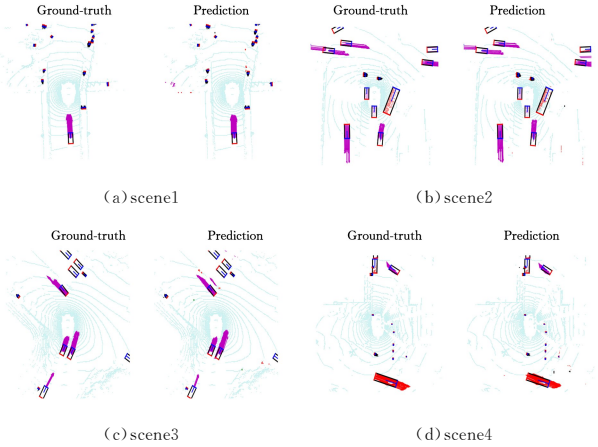


图 5　不同场景下真值轨迹和预测轨迹的对比结果
(电子版为彩图)

Fig. 5　Comparison results of true trajectory and predicted trajectory in different scenarios

4.3　消融实验

本节比较了各种加权邻接矩阵核函数对性能的影响。其中，对加权邻接矩阵 $\boldsymbol{A}_t$ 顶点的贡献进行加权，核函数是交通参与者之间社会关系的先验知识。在核函数的选择上，使用式(10)中定义的 L2 范数测量的交通参与者之间的距离模拟它们对彼此的影响。然而，这与我们所观察到的实际情况不一样，因为交通参与者更容易受到与其距离近的其他相似物体的影响。因此，本研究使用交通参与者的相似性度量来解决这一问题。我们采用了 L2 范数的倒数，如式(11)定义的那样来度量，并且将 ϵ 加到分母中，确保分母不为 0。式(12)中定义的函数是高斯径向基函数[31]。这些核函数的性能通过消融实验表示在表 3 中。我们将 1 设为加权邻接矩阵中的基线，通过表 3 列出的结果可以看到性能最好的是式(1)中定义的 $a^{ij}_{\mathrm{sim},t}$，其与式(11)的不同之处在于条件 $\| v_t^i - v_t^j \|_2 = 0$，式(1)中设定当 $\| v_t^i - v_t^j \|_2 = 0$ 时，$a^{ij}_{\mathrm{sim},t} = 0$。这表明，两个交通参与者在同一位置时，会被认为是同一个物体。如果没有这个条件，模型中交通参与者之间的关系无法正确表示出来。因此，本文在实验中使用 $a^{ij}_{\mathrm{sim},t}$ 来定义邻接矩阵。

$$a^{ij}_{L_2,t} = \| v_t^i - v_t^j \|_2 \tag{10}$$

$$a^{ij}_{\mathrm{sim}_\epsilon,t} = \frac{1}{\| v_t^i - v_t^j \|_2 + \epsilon} \tag{11}$$

$$a^{ij}_{\exp,t} = \frac{\exp(-\| v_t^i - v_t^j \|_2)}{\sigma} \tag{12}$$

表 3　不同核函数的邻接矩阵对误差的影响

Table 3　Influence of adjacency matrix of different kernel functions on error

Kernel function	static	velocity≤5 m/s	velocity>5 m/s
$a^{ij}_{L_2,t}$	0.3723	0.4109	1.7890
$a^{ij}_{\mathrm{sim}_\epsilon,t}$	0.3778	0.3787	1.7654
$a^{ij}_{\exp,t}$	0.3906	0.4235	1.8765
1	0.3820	0.3932	1.5776
$a^{ij}_{\mathrm{sim},t}$	0.3450	0.2720	1.1074

此外，本文针对点云输入帧数对实验结果的影响进行了分析。当输入网络的 BEV 点云图的帧数增多时，其时间复杂度和空间复杂度都会显著增加，因此本文需要在输入帧数和模型性能之间找到平衡点。如图 6 所示，当帧数超过 5 时，模型精度饱和，性能增益较小，因此我们在实验中每次输入 5 帧。

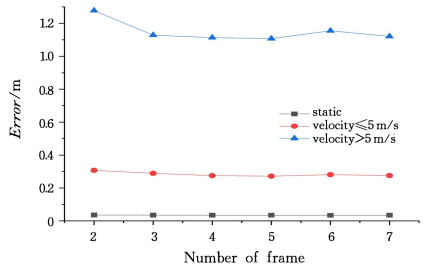


图 6　输入帧数对性能的影响

Fig. 6　Performance impact on number of input frames

结束语　本文提出了交互特征与场景特征融合的方法来进行轨迹预测。该方法一方面通过时空图卷积网络对交通参与者之间的交互模式进行编码，来表达交互关系；另一方面采用时空金字塔网络对周围信息进行场景建模，获取场景特征。本文通过不同特征网络的融合和适配来实现端到端的方法。轨迹预测的实验表明，在速度＜5m/s 和速度＞5m/s 的场景下，本文提出的方法的轨迹误差小于对比方法。同时在分类精度的实验结果中，本文方法和对比方法相比，平均分类精度和总体网格分类精度均有所提升。对于轨迹预测问题的后续研究，我们会从传感器损伤、复杂交通环境和精密地图提供的交通规则等方面入手。

参 考 文 献

[1] MOZAFFARI S, AL-JARRAH O Y, DIANATI M, et al. Deep Learning-based Vehicle Behaviour Prediction for Autonomous Driving Applications: A Review[J]. arXiv: 1912.11676, 2019.

[2] LEE N, CHOI W, VERNAZA P, et al. DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu, IEEE, 2017: 2165-2174.

[3] ZENG W L, CHEN Y H, YAO R Y, et al. Application of Spatial-Temporal Graph Attention Networks in Trajectory Prediction for Vehicles at Intersections[J]. Computer Science, 2021, 48(S1): 334-341.

[4] LI L H, ZHOU B, LIAN J, et al. Research on pedestrian trajectory prediction method based on social attention mechanism[J]. Journal on Communications, 2020, 41(12): 175-183.

[5] JUSTS D J, NOVICKIS R, OZOLS K, et al. Bird's-eye view image acquisition from simulated scenes using geometric inverse perspective mapping[C]// 2020 17th Biennial Baltic Electronics

Conference(BEC). Tallinn, 2020: 1-6.
[6] CHEN S, LIU B, FENG C, et al. 3D Point Cloud Processing and Learning for Autonomous Driving[J]. IEEE Signal Processing Magazine, 2021, 38(1): 68-86.
[7] LI B L, YANG D, WANG L, et al. Weak Echo Signal Processing of 1 550 nm Coherent Laser Wind Radar[J]. Piezoelectrics and Acoustooptics, 2022, 44(2): 333-338.
[8] LEFÉVRE S, VASQUEZ D, LAUGIER C. A survey on motion prediction and risk assessment for intelligent vehicles[J]. Robomech Journal, 2014, 1(1): 1-14.
[9] YOU L, HAN X W, HE Z W, et al. Improved Sequence-to-Sequence Model for Short-term Vessel Trajectory Prediction Using AIS Data Streams[J]. Computer Science, 2020, 47(9): 169-174.
[10] ZHOU Y, TUZEL O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City, IEEE, 2018: 4490-4499.
[11] LUO W, YANG B, URTASUN R. Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City, IEEE, 2018: 3569-3577.
[12] LEFEVRE S, VASQUEZ D, LAUGIER C. A survey on motion prediction and risk assessment for intelligent vehicles[J]. ROBOMECH Journal, 2014, 1(1): 1-9.
[13] SHI S, WANG X, LI H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, IEEE, 2019: 770-779.
[14] ZENG W Y, WANG S L, LIAO R J, et al. DSDNet: Deep Structured self-Driving Network[C]// 2020 European Conference Computer Vision(ECCV). Glasgow, 2020: 156-172.
[15] SCHREIBER M, HOERMANN S, DIETMAYER K. Long-Term Occupancy Grid Prediction Using Recurrent Neural Networks[C]// 2019 International Conference on Robotics and Automation(ICRA). Montreal, 2019: 9299-9305.
[16] CHEN X, MA H, WAN J, et al. Multi-View 3D Object Detection Network for Autonomous Driving[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 6526-6534.
[17] YUAN Z, SONG X, BAI L, et al. Temporal-Channel Transformer for 3D Lidar-Based Video Object Detection for Autonomous Driving[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(4): 2068-2078.
[18] CUI H, RADOSAVLJEVIC V, CHOU F C, et al. Multimodal Trajectory Predictions for Autonomous Driving using Deep Convolutional Networks[C]// 2019 International Conference on Robotics and Automation(ICRA). 2019: 2090-2096.
[19] XU J, XIAO L, ZHAO D, et al. Trajectory Prediction for Autonomous Driving with Topometric Map[J]. arXiv: 2105. 03869, 2021.
[20] ZENG W Y, LUO W J, SUO S, et al. End-to-end Interpretable Neural Motion Planner[C]// 2019 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2019: 8660-8669.
[21] CASAS S, LUO W J, URTASUN R. IntentNet: Learning to Predict Intention from Raw Sensor Data[C]// CoRL 2018. 2018: 947-956.
[22] ZHANG Z S, GAO J Y, MAO J H, et al. STINet: Spatio-Temporal-Interactive Network for Pedestrian Detection and Trajectory Prediction[C]// 2020 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2020: 11346-11355.
[23] TRAN D, BOURDEV L, FERGUS R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// IEEE International Conference on Computer Vision. IEEE, 2015: 4489-4497.
[24] WU Z, PAN S, CHEN F, et al. A Comprehensive Survey on Graph Neural Networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 32(1): 4-24.
[25] MARINO K, SALAKHUTDINOV R, GUPTA A. The More You Know: Using Knowledge Graphs for Image Classification [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu, IEEE, 2017: 20-28.
[26] SHEN Y, LI H, YI S, et al. Person Re-identification with Deep Similarity-Guided Graph Neural Network[C]// European Conference on Computer Vision. Cham: Springer, 2018: 508-526.
[27] WU P, CHEN S, METAXAS D. MotionNet: Joint Perception and Motion Prediction for Autonomous Driving Based on Bird's Eye View Maps[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, IEEE, 2020: 11382-11392.
[28] LIU X, QI C R, GUIBAS L J. FlowNet3D: Learning Scene Flow in 3D Point Clouds[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, IEEE, 2019: 529-537.
[29] GU X, WANG Y, WU C, et al. HPLFlowNet: Hierarchical Permutohedral Lattice FlowNet for Scene Flow Estimation on Large-Scale Point Clouds[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, IEEE, 2019: 3254-3263.
[30] SCHREIBER M, HOERMANN S, DIETMAYER K. Long-Term Occupancy Grid Prediction Using Recurrent Neural Networks[C]// 2019 International Conference on Robotics and Automation(ICRA). Montreal, 2019: 9299-9305.
[31] SCHÖLKOPF B, TSUDA K, VERT J. A Primer on Kernel Methods[M]. Massachusetts: MIT Press, 2004.

FANG Yang, born in 1991, Ph.D, lecturer, is a member of China Computer Federation. His main research interests include computer vision and pattern recognition, visual object tracking, lidar-based 3D sensing and perception for AD system.

ZHAO Ting, born in 1995, postgraduate. Her main research interests include big data, lidar sensing and trajectory prediction.

(责任编辑:柯颖)