基于时空Transformer特征融合的车辆轨迹预测

赵文红[1]，王巍[2]，万子璐[3]
（1.嘉兴南湖学院公共基础教学部，浙江 嘉兴 314001；2.电磁空间安全全国重点实验室，浙江 嘉兴 314033；
3.浙江工业大学信息工程学院，浙江 杭州 310013）

摘　要：在复杂的交通环境下，自动驾驶汽车需要充分地分析周围交通物体的运动方向、运动速度等信息，并准确预测未来的轨迹。针对这个问题，提出了一种基于时空Transformer的网络模型。该模型首先利用空间自注意力机制，通过捕捉同一时刻下车辆间的空间相互作用，实现对多车空间关系交互性的精确建模；随后通过时间自注意力机制提取连续帧的时间依赖关系，以此生成一组能够反映车辆动态行为的时空特征；最后这些特征被送入解码器，以预测所有车辆在未来5 s内的运动轨迹。在公开的NGSIM数据集上进行了训练和验证，与其他的先进方案相比，该模型在未来5 s的轨迹预测中具有更高的准确性和精度，长期预测准确率比先进方案提高14.6%。

关键词：自动驾驶；轨迹预测；多车交互；Transformer
中图分类号：TP183
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024192

Vehicle trajectory prediction based on spatio-temporal Transformer feature fusion

ZHAO Wenhong[1], WANG Wei[2], WAN Zilu[3]
1. Department of Public Basic Education, Jiaxing Nanhu University, Jiaxing 314001, China
2. National Key Laboratory of Electromagnetic Space Security, Jiaxing 314033, China
3. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310013, China

Abstract: In complex traffic environments, autonomous vehicles must thoroughly analyze the motion direction, speed, and other information of surrounding traffic objects to accurately predict future trajectories. A network model based on spatio-temporal Transformer was proposed to address this issue. The framework initially employs a spatial self-attention mechanism to capture the spatial interactions between vehicles at the same moment, achieving precise modeling of the spatial relationship interactivity among multiple vehicles. Subsequently, a temporal self-attention mechanism was utilized to extract the temporal dependencies between consecutive frames, thereby generating a set of spatiotemporal features that reflect the dynamic behavior of vehicles. These features were then fed into a decoder to predict the motion trajectories of vehicles over the next 5 s. The proposed model was trained and validated on the publicly available NGSIM dataset. Compared to other state-of-the-art schemes, our scheme demonstrates greater accuracy and precision in trajectory prediction over the subsequent 5 s. The long-term forecasting accuracy is increased by 14.6% compared to the advanced schemes.

Keywords: autonomous driving, trajectory prediction, multi-vehicle interaction, Transformer

收稿日期：2024-06-28；**修回日期：**2024-10-17
基金项目：国家自然科学基金资助项目（No.62231027, No.U19B2015, No.U21B2001）；嘉兴南湖学院科研基金资助项目（No.62211ZL）
Foundation Items: The National Natural Science Foundation of China (No. 62231027, No. U19B2015, No. U21B2001), Research Fund Project of Jiaxing Nanhu University (No.62211ZL)

0 引言

轨迹预测作为规划的一部分，可以很好地反映周围移动实体的未来行为，在感知和决策之间架起桥梁。车辆轨迹预测是指根据历史和实时的交通数据，预测未来一段时间内各个车辆（如轿车、卡车、摩托车等）在道路网络上的运动状态和位置。近年来，随着深度学习技术的发展，越来越多的研究者尝试使用神经网络模型来解决车辆轨迹预测问题，比如卷积神经网络（CNN）、长短期记忆网络（LSTM）、图神经网络（GNN）在轨迹预测的研究中已有了广泛的应用。近年来，Transformer模型作为一种强大的序列建模工具，受到了广泛关注。Transformer模型通过自注意力机制可以有效地捕捉序列中的长距离依赖关系，并且具有并行计算和易于扩展的优势。

CNN可以提取图像中的局部特征，利用池化层大幅度降低参数的数量级，通过全连接层输出想要结果。CNN常应用于图片分类与检索、目标定位检测、目标分割等，目前在车辆轨迹预测中也有较多的应用，如Luo等[1]提出了一种用于快速目标检测、跟踪和运动预测的卷积网络模型。该模型将鸟瞰激光雷达数据作为输入，处理跨越空间和时间的三维卷积，然后添加2个额外的卷积层分支：一个分支计算在给定位置成为车辆的概率，另一个分支预测当前帧以及未来几帧的边界框。他们认为这样的结构可以预测运动，因为该模型可以从多个帧的输入中学习速度和加速度特征。但是，预测分支简单地将三维卷积特征图作为输入，因此所有物体的视觉特征都表示在同一个特征图中，这导致模型失去了对目标物体的跟踪，在拥挤的场景中不能很好地执行。

LSTM可以保留较长序列数据中的“重要信息”，忽略不重要的信息，被广泛应用于时间特征的提取和建模。在车辆行为预测中，LSTM是很常用的模型。温惠英等[2]设计的基于生成对抗网络的轨迹预测模型，该模型采用了LSTM的编码器-译码器结构，通过输入给定的历史换道轨迹，经解码器生成预测时段环道轨迹；Khosroshahi等[3]和Phillips等[4]使用LSTM对十字路口的车辆机动进行分类；Kim等[5]提出了一种基于LSTM的高效车辆轨迹预测模型，该模型在未来的0.5 s、1 s和2 s时间间隔内预测车辆在占用网格中的位置；Lee等[6]提出了一种结合了条件变分自动编码器（CVAE）和循环神经网络（RNN）编码器-解码器的轨迹预测模型，虽然其允许通过对CVAE采样进行多模态预测，但该模型只能提供来自预测分布的样本，而不能提供分布本身的估计；蔡英凤等[7]提出了一种非均匀步长的时间序列数据划分方法，将属于特定行为的车辆时序信息进行分类，以LSTM为基本的神经网络框架，用注意力机制判断输入时序信息中各个时间步信息的重要程度，分配不同的权重值，然后以目标车辆及其周边车辆的历史轨迹信息作为算法输入，用来预测目标车辆将来的运动行为；Alahi等[8]提出了一种Social-LSTM模型，该模型通过将局部静态场景图像作为LSTM的额外输入，可以在框架中对人-空间交互进行建模，这可以在同一框架下对人与人以及人与空间的互动共同建模。这使得其可以自动学习在时间重合的轨迹之间发生的典型交互，无须任何其他注释就可以学习行人在交通环境所遵循的规则和常识；Gupta等[9]提出的社交生成式对抗网络（CS-LSTM）结合了一种新颖的池化机制和生成对抗网络，从本质上解决行人轨迹的多模态；Zhang等[10]提出的状态优化长短期记忆网络（SR-LSTM）引入了一种消息传递和选择机制，以捕获邻居的关键当前意图。

然而基于LSTM的方法使用一个内存有限且单一的向量来记忆历史信息，并且通常难以处理复杂的时间依赖性。因此，池化机制、注意力机制和图卷积方法被用来模拟空间的相互作用。Deo等[11-12]提出了高速公路上周边车辆机动分类和运动预测的统一框架。首先，使用LSTM模型表示所有观察到的车辆（被预测的车辆及其附近车辆）的轨道历史信息和相对位置，作为上下文向量。然后，利用该上下文向量进行机动分类，利用另一个LSTM模型预测飞行器的未来位置。考虑到LSTM模型无法捕捉场景中所有汽车运动的相互依赖性，他们随后通过在文献[13]中添加卷积社会池化层来增强方案。该改进模型可以获取周围物体的运动状态及其空间关系，从而提高未来运动预测的准确性。然而，所有这些模型每次只能预测一辆特定汽车的轨迹。因此，这些现有的方法如果要预测所有周围物体的轨迹，需要大量的计算能力，这是非常低效的。Ivanovic等[14]提出了一种图形结构模型，可以在高动态和多模态场景中同时预测多个行人的许多潜在未

来轨迹，这是一种新颖的多智能体建模方法，它明确地解释了人类行为的关键，即它们是多模态的、动态可变的。Zhao 等[15]提出了一种新的基于图的信息共享网络（GISNet），该网络允许目标车辆与周围车辆之间的信息共享，但其没有考虑每辆车的时间信息和历史轨迹。Li 等[16]提出了一种名为图交互感知（GRIP）的新方案，旨在有效地预测自动驾驶汽车周围交通代理的轨迹。GRIP 使用图形来表示近距离对象的交互作用，用多个图形卷积来提取特征，然后使用 LSTM 模型进行预测。Xu 等[17]提出了一种自适应参数矩阵来协调和优化全局时空图，采用堆叠图卷积模块提取车辆历史轨迹数据的全局时空特征。Xu 等[18]提出了一种多视图自适应层次化空间图卷积网络，通过构建多视图逻辑网络结合图卷积和区域聚类技术，预测异构交通参与者的未来轨迹。这些方法的局限性在于只模拟了近距离的交通主体之间的相互作用，而忽略了超出给定空间界限的交通主体的影响。此外，这些方法大多针对单一场景中的轨迹预测，在处理密集且复杂的城市环境中可能有很大的局限性。

由于 Transformer[19]独特的注意力机制和在自然语言处理（NLP）中展现的优越性能，近年来将 Transformer 架构应用于轨迹预测任务的兴趣越来越强烈。Transformer 模型之所以适合车辆轨迹预测任务，主要得益于其独特的架构设计和处理序列数据的能力，比如通过自注意力机制，能够有效捕捉序列数据中的长距离依赖关系，这对于理解车辆运动的长期模式至关重要。Giuliari 等[20]首次提出采用 NLP 来进行轨迹预测，提出了一个多智能体框架，其中每个人都由 Transformer 网络的一个实例建模，每个 Transformer 网络根据个体之前的运动来预测其未来的运动，讨论了 TF（Transformer network）和 BERT（bidirectional Transformer）2 个模型，分析行人轨迹预测，研究数据表明，Transformer 不仅能很好地预测行人的轨迹，还显示出了更好的长期预测能力，能预测合理的多个未来轨迹。Yu 等[21]提出了一个基于注意力机制的时空图变换器网络框架，用于预测行人的运动轨迹，其中时间变换器独立地处理每个行人的轨迹嵌入，并输出具有时间依赖性的更新嵌入，空间变换器使用一种新的基于 Transformer 的图卷积机制来提取行人之间的空间交互，它将人群建模为图，并学习通过空间变换器和时间变换器的时空交互。Liu 等[22]提出了一个名为 mmTransformer 的新型多模态运动预测框架，采用堆叠的 Transformer 网络构架，开发了一种新颖的训练策略，通过将轨迹提议划分为不同的空间集群，基于真实轨迹终点的空间分布来分别聚合历史轨迹、道路信息以及交互信息，实现多模态预测。以上方法的局限性在于，只针对那些物理位置上彼此接近的交通参与者之间的交互信息，即主要是基于局部空间的交互信息，而不是整个交互网络中的所有交通参与者的交互信息。在低速情况下，邻近车辆或行人可能是影响当前运动轨迹的主要因素，但在高速情况下，更远距离的车辆也可能因为速度变化而对当前车辆产生影响。Ngiam 等[23]提出了一种联合预测模型，使用注意力机制和序列掩码策略来预测自动驾驶环境中所有智能体的行为和交互。景荣荣等[24]将交通参与者的历史轨迹和周围交通环境信息编码为多通道图，并作为模型输入，利用改进的 Transformer 对交通环境进行建模，捕捉交通智能体与交通环境之间的交互信息，以预测未来运动轨迹。然而这种方法只能对单智能体进行轨迹预测，不能同时对多个智能体进行预测。王庆荣等[25]设计了一种结合门控循环单元（GRU）和 Transformer 的模型结构，使用 GRU 从输入的车辆历史轨迹序列中提取时间序列特征，利用具有双层多头注意力机制的 Transformer 编码器来捕获车辆间时空交互特征。虽然分别获取了车辆的时间和空间交互特征，但将时间和空间交互分开处理可能导致模型无法充分学习到它们之间的内在联系和相互影响。

本文致力于研究在高速公路的交通环境下，如何利用历史轨迹以及特征数据，预测群体车辆在未来一段时间内的位置和运动状态。针对前述的国内外已有的工作以及在轨迹预测问题中存在的局限性，现提出一种基于时空 Transformer 的网络模型，以预测车辆的未来轨迹。具体而言，在空间维度上，应用空间自注意力机制来捕捉路网中所有车辆之间的交互信息；在时间维度上，利用时间自注意力机制来独立地提炼每一车辆的时间特征。通过自回归分析，综合空间特征和时间特征生成所有车辆的预测轨迹。研究结果通过美国高速公路 NGSIM 数据集验证，证实了所提模型的有效性。

1　问题描述

1.1　参考坐标系

场景采用固定参考坐标系，即笛卡儿坐标系，如图1所示，其中，x轴指向与前进方向垂直的方向，y轴指向前进方向，单位均为米。

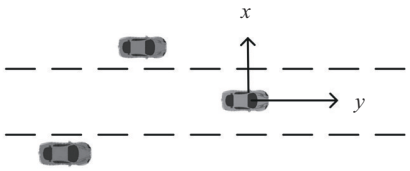


图1　参考坐标系

1.2　输入与输出表示

车辆轨迹预测旨在通过精确分析历史车辆状态，提供高准确度的未来轨迹预测，以支持安全和高效的驾驶决策。在真实场景中，模型不只是对单个车辆进行轨迹预测，而是对可观察到的车辆进行实时预测。因此，群体车辆的输入为

$$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t] \tag{1}$$

其中，

$$\begin{aligned}\boldsymbol{x}_i=\{&(x_{i0},y_{i0},l_{i0},w_{i0},v_{i0},\tau_{i0}),\cdots,\\&(x_{\mathrm{in}},y_{\mathrm{in}},l_{\mathrm{in}},w_{\mathrm{in}},v_{\mathrm{in}},\tau_{\mathrm{in}})|i\in(1,t)\}\end{aligned} \tag{2}$$

是某一时刻t下所有车辆的历史特征向量，x和y为车辆的坐标，l和w为车辆的长度和宽度，v代表车辆的速度，τ代表车辆的类型（轿车、卡车、摩托车）。

由此，输入的历史轨迹等信息可表示为

$$F_{\mathrm{input}}=\mathbf{R}^{T\times N\times D} \tag{3}$$

其中，$\mathbf{R}$为实数集，T为历史时间，N为最大可见车辆的数量，D为历史特征向量的维度。

预测的是未来3 s，即15帧的所有车辆的轨迹。模型输出为

$$Y=[y_{t_0},y_{t_1},\cdots,y_{t_f}] \tag{4}$$

其中，

$$y_i=\{(x_{i0},y_{i0},x_{i1},y_{\mathrm{i1}},\cdots,x_{\mathrm{in}},y_{\mathrm{in}})|i\in(t_0,t_f)\} \tag{5}$$

是某一未来时刻下，所有被预测的车辆的未来轨迹。

2　轨迹实时预测网络模型

时空Transformer网络模型如图2所示，主要由3个部分组成：输入特征表示、时空Transformer编码器和时空Transformer解码器。首先，通过预处理车辆的历史特征数据，构建了一个时空矩阵，并将其映射到更高维的特征空间中。在编码阶段，采用了空间自注意力机制，该机制能够精确捕获车辆在同一时间内的空间相互作用。此外，还引入了时间自注意力机制，以捕捉每一车辆在连续时间帧内的时间依赖关系。这些机制的结合，通过可分离卷积进一步提炼出精细的时空特征。在解码阶段，本文设计的时空Transformer解码器不仅利用编码器提供的时空特征，还结合了先前的预测坐标，以精细化输出嵌入。特别地，引入了时间掩码自注意力技术，该技术确保了模型在计算注意力得分时仅考虑历史信息，

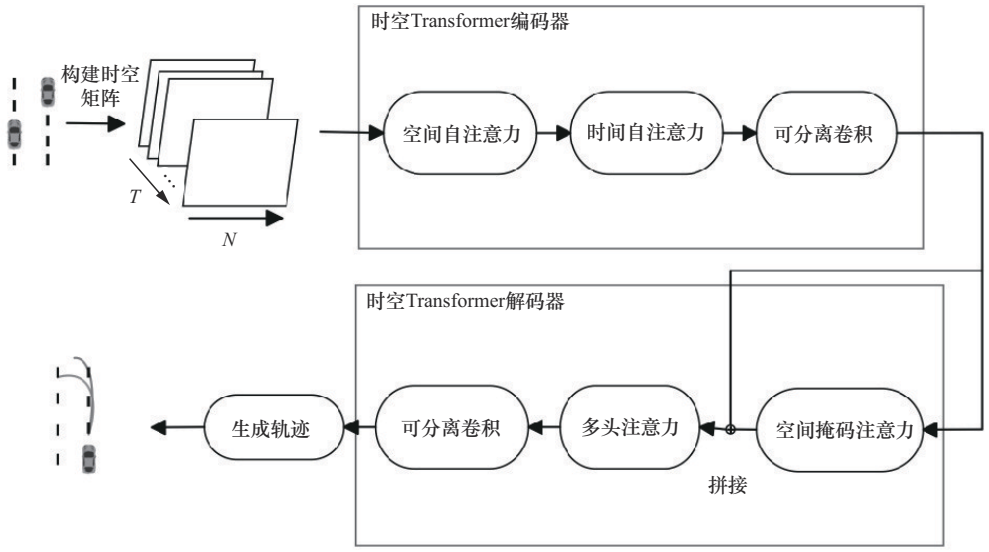


图2　时空Transformer网络模型

从而提高了预测的准确性和可靠性。最后，通过轨迹生成器，该模型能够输出所有车辆的未来运动轨迹。该模型在处理复杂的交通场景时，展现了卓越的性能，特别是在捕捉车辆间的复杂交互和预测长期轨迹方面。

2.1　网络输入的时空矩阵构建

在对数据集进行预处理后，分别在时间和空间上建立矩阵，然后组成在N个车辆和T帧下的时空图$\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$。其中，$\boldsymbol{V}=\{x_{it}|i\in(1,N),t\in(1,T)\}$表示节点集，包含了所有车辆的特征向量；边集$E$则有2个子集，分别表示同一帧内不同车辆之间的联系以及连续帧中同一车辆的联系边，前者空间图记为$\boldsymbol{E}_s=\{(x_{it},x_{jt}|i,j\in(1,N),t\in(1,T))\}$，后者时间图记为$\boldsymbol{E}_T=\{(x_{it_0},x_{it_1})|i\in(1,N),t_0,t_1\in(1,T)\}$。2种边集构建方法具体如下。

1) 空间图构建。为提升空间自注意力机制在计算注意力得分时的效率，采用图的方式表示车辆间的相互作用。首先通过建立一个三维的布尔类型矩阵来编码任意车辆在时刻t的连接状态，即在某时刻t下，首先计算每个车辆的中心位置。

$$d_{ij}=\sqrt{(x_{t_i}-x_{t_j})^2+(y_{t_i}-y_{t_j})^2} \tag{6}$$

其中，d_{ij}表示节点i与节点j之间的欧氏距离。

然后判断2个车辆是否邻接。若它们之间的中心距离小于或等于某个阈值k，则强调它们之间的相互作用，并将边联系关系A_{ij}设为正确，否则设置为错误。

$$A_{ij}=\begin{cases}\text{正确},d_{ij}\leqslant k\\ \text{错误},d_{ij}>k\end{cases} \tag{7}$$

其中，k为距离阈值。

空间图如图3所示，这个三维矩阵代表了时空图的建立，其中，T代表历史时间，N代表最大可见车辆数量，A_{ij}代表某个时刻下2个车辆是否邻接。

2) 时间图构建。为确保时间掩码自注意力机制仅考虑历史时间的车辆及其特征，从而有效地计算注意力得分，引入了信息屏蔽技术。具体实现包括建立一个二维的布尔类型矩阵，在历史时刻t下，若某个车辆在研究范围内，则将边联系关系设置为正确，否则设置为错误。时间图如图4所示，A_{it}代表t时刻下节点i是否在研究范围内。

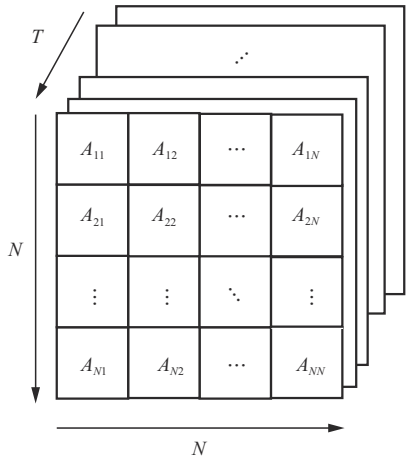


图3　空间图

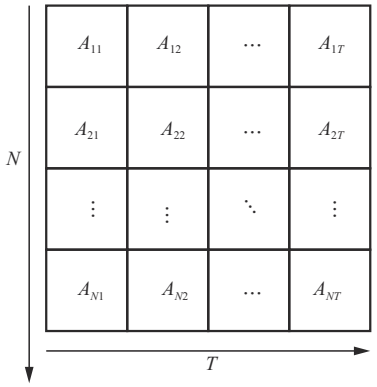


图4　时间图

2.2　时空Transformer编码器

时空Transformer编码器主要由空间自注意力、时间自注意力和可分离卷积组成，并插入残差连接和层归一化，帮助信息在网络中更好地传递和流动，提高网络的鲁棒性和学习能力。

1)空间自注意力。如图5所示，对于时刻t场景的每个节点i，都可以通过线性投影计算出它们的查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$。

$$\boldsymbol{Q}_i^t=\boldsymbol{W}_q h_i^t \tag{8}$$

$$\boldsymbol{K}_i^t=\boldsymbol{W}_k h_i^t \tag{9}$$

$$\boldsymbol{V}_i^t=\boldsymbol{W}_v h_i^t \tag{10}$$

其中，h_i^t为输入嵌入，$\boldsymbol{W}_q$为查询权重矩阵，$\boldsymbol{W}_k$为键权重矩阵，$\boldsymbol{W}_v$为值权重矩阵。

然后，求得节点i和节点j之间的空间注意力得分

$$m_{ij}^t=\boldsymbol{Q}_i^t\boldsymbol{K}_j^t \tag{11}$$

再将所有节点j到节点i的注意力得分在空间边的权重上归一化并求和得到节点i的单个注意头

$$\text{head}_i^t = \text{softmax}\left(\frac{m_{ij}^t}{\sqrt{d_K}}\right)\boldsymbol{V}_j \tag{12}$$

其中，d_K为键向量的维度。

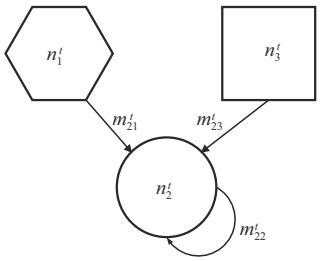


图5　t时刻下的空间自注意力示意

重复以上提取单个注意头的过程，将这些单个注意头连接，并投射到具有完全连接层的输出嵌入中。

$$\text{MultiHead}_i^t = \boldsymbol{W}_o \cdot \text{concat}([\text{head}_{i_0}^t,\cdots,\text{head}_{i_h}^t]) \tag{13}$$

其中，$\boldsymbol{W}_o$为权重矩阵，concat表示直接拼接。

空间多头注意力考虑了车辆之间的欧氏距离，也处理了复杂场景下车辆的交互性，因此能够清晰地反映车辆之间的逻辑关系。

2) 时间自注意力。如图6所示，时间自注意力计算了不同时刻下同一车辆的时间注意力得分。与空间自注意力类似，时间自注意力用到了多头注意力，其不同点在于计算的是同一个节点在不同时间下的注意力得分，如式(14)所示。

$$\text{head}_i = \text{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V}_i \tag{14}$$

然后计算其多头注意力得分为

$$\text{MultiHead}_i = \boldsymbol{W}_u \cdot \text{concat}([\text{head}_{i_0},\cdots,\text{head}_{i_h}]) \tag{15}$$

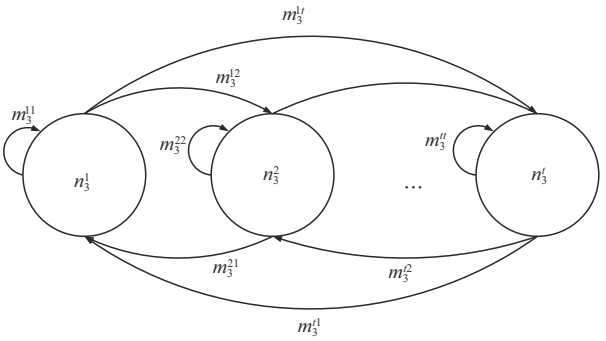


图6　同一车辆在不同时间下的自注意力示意

3) 可分离卷积。对比全连接网络，可分离卷积[16]有着更好的感受野效果和泛化能力，更少的参数量和计算量，故选择可分离卷积代替全连接网络。

交通场景的复杂性和流通性，使得交通参与者之间不可避免地形成了紧密的时空联系，为了深入挖掘交通参与者的时间特性和空间特性，并准确捕捉路网的时空属性，本文方法将需要融合的特征直接拼接。

2.3　时空Transformer解码器

为使得将时空Transformer编码器输出轨迹的相对位置信息有效传递到时空Transformer解码器中，需要对输出嵌入添加位置编码。

$$\text{PE}_{(\text{pos},2i)} = \sin\left(\frac{\text{pos}}{10\,000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{16}$$

$$\text{PE}_{(\text{pos},2i)} = \cos\left(\frac{\text{pos}}{10\,000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{17}$$

其中，pos是位置信息，i是维度，d_{model}是output embeddings的总维度。

对比时空Transformer编码器中的空间自注意力和时间自注意力，时空Transformer解码器使用了时空掩码自注意力来确保只能考虑时刻t之前的历史时间，使得预测结果仅依赖已生成的输出轨迹。与时空Transformer编码器相同，采用了可分离卷积，并在时空掩码自注意力与可分离卷积之间插入了多头注意力，对时空Transformer编码器的输出进行多头注意力计算。

为了最小化预测轨迹与真实轨迹之间的差值，采用L2-loss函数。

$$\text{Loss} = \sum_{t=t_1}^{T}\left|Y_{\text{pred}}^t - Y_{\text{GT}}^t\right|^2 \tag{18}$$

其中，Y_{pred}^t和Y_{GT}^t分别表示t时刻下车辆的预测位置和真实位置。

3　实验与分析

3.1　数据预处理

模型在美国公共数据集NGSIM上进行了评估，NGSIM数据集的收集是通过设置视频摄像头和传感器来实现的，涵盖了美国加利福尼亚州南行的详细车辆轨迹信息US-101数据集和旧金山湾区东行的详细车辆（包括轿车、摩托车、卡车）轨迹信息I-80数据集等，US-101数据集的研究区域如图7所

示，研究长度为2 100 ft（即640.08 m），包括5条主要车道。图8为I-80数据集的研究区域，其研究长度为1 650 ft（即502.92 m），有6个车道和2个汇接出入口。

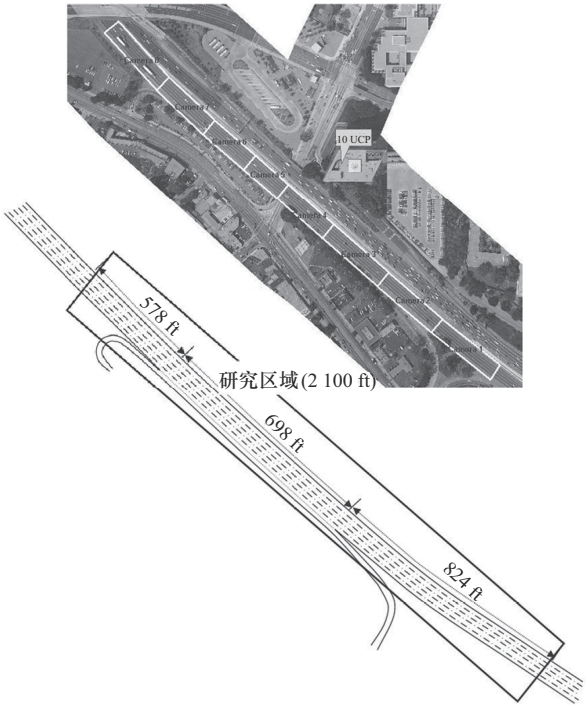


图7　US-101数据集的研究区域

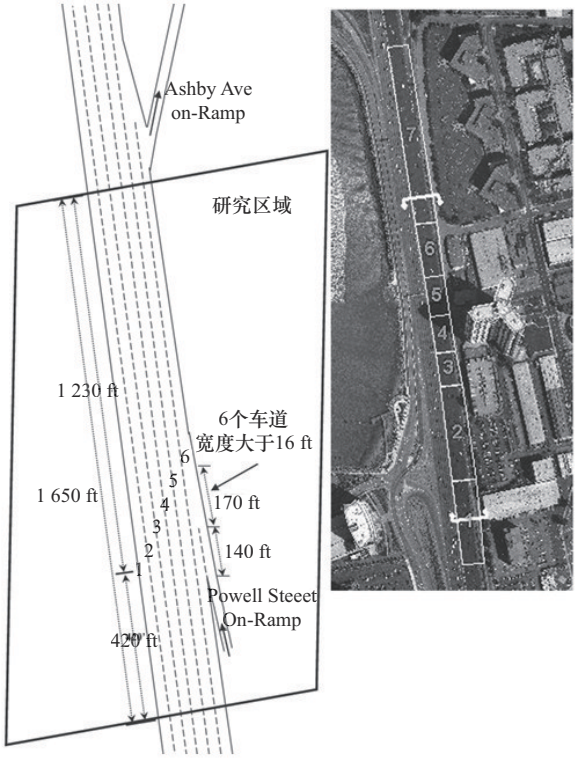


图8　I-80数据集的研究区域

US-101和I-80数据集由6个15 min的子集组成，每个子集的20%组成测试集，其他数据则划分为训练集和验证集，并采用留出法从训练集中分离出20%作为验证集。为了与文献[6]中的实验保持一致，采用了相同的降采样频率，即降采样2倍。此外，为了更好地体现不同车辆的异质性，网络输入特征除了车辆ID、帧ID和坐标以外，还额外选取了车辆的长度、宽度、速度、类型信息。所有长度单位均由英尺转换为米，以统一数据格式。

3.2　实验实施细节

实验在一台运行Ubuntu 16.04操作系统的台式计算机上执行，该计算机配备了2.30 GHz Intel（R） Xeon（R） Gold 5218 CPU、32 GB内存和NVIDIA Tesla V100-SXM2显卡。实验基于PyTorch框架实现，采用6层堆叠的时空Transformer编码器和解码器。模型优化采用Adam算法，设置学习率为0.001。

实验采用均方根误差（RMSE）来计算历史轨迹误差，计算式为

$$\mathrm{RMSE}^t=\frac{\sqrt{\sum_{i=1}^{N}[(x_{ti}-x'_{ti})^2+(y_{ti}-y'_{ti})^2]}}{N} \tag{19}$$

其中，x_{ti}和y_{ti}表示在t时刻下预测的编号为i的车辆的轨迹坐标，x'_{ti}和y'_{ti}表示t时刻下预测的编号为i的车辆的真实的轨迹坐标，N代表t时刻下所预测轨迹的车辆的总数，RMSE^t表示t时刻下的指数均方误差。

用平均位移误差（ADE）和最终位移误差（FDE）2个指标来评估模型最终的性能，其中，ADE为预测时间内预测位置与真实位置的平均欧氏距离，FDE为ADE的最后一项。显然，ADE显示了平均的预测性能，FDE则反映了端点的预测精度。

$$\mathrm{ADE}=\frac{\sum_{t=1}^{T}\mathrm{RMSE}^t}{T} \tag{20}$$

$$\mathrm{FDE}=\mathrm{RMSE}^T \tag{21}$$

其中，T表示未来可预测的最大时间间隔。

3.3　实验结果对比与分析

本节引入了几种基线方法及一些先进方法，并与基于时空Transformer的网络模型进行比较，验证所提模型的有效性。

等速卡尔曼滤波（CV）：一种用等速卡尔曼滤波来预测未来轨迹的基线方法。

原始长短记忆网络（V-LSTM）：一种基于LSTM编码器−解码器框架的基线方法。

生成式对抗网络GAIL-GRU[26]：一种用生成式对抗学习模型来预测车辆未来轨迹的方法。

社会池化−长短期记忆网络（CS-LSTM）[13]：使用卷积社会池化层，从车辆的历史轨迹中提取特征，实现了推断相互依赖的相邻车辆的运动。为了更好地满足当前自动驾驶领域的需求，基于时空Transformer的网络模型实现了对所有车辆未来轨迹的实时预测，而其他模型则只是预测了目标车辆的未来轨迹，未能充分考虑对周围车辆的交互感知。对比结果如表1所示，所提模型无论是在哪个预测时域上的表现都优于其他模型，特别是在长时间预测上有更大的优势，与CV和V-LSTM相比，ADE分别提高了38.3%和33.9%，FDE分别提高了44.2%和40.5%；与GAIL-GRU和CS-LATM相比，ADE分别提高了19.5%和7.5%，FDE分别提高了20.8%和14.6%。实验数据表明，所提模型在轨迹预测的准确性方面具有明显优势。

表1 基于时空Transformer的网络模型与其他模型在US-101和I-80数据集下的比较

模型	ADE	FDE
CV	3.42	6.68
V-LSTM	3.19	6.27
GAIL-GRU	2.62	4.71
CS-LSTM	2.28	4.37
时空Transformer	2.11	3.73

3.4 模型参数研究

根据1.1节的讨论，若两辆车的中心距离小于或等于阈值k，则强调它们之间的空间关系，反之则不强调。实验选择了5、15、25、35 m作为邻接阈值k的不同取值，并进行了相应的实验，结果如图9所示。当阈值k为5 m和10 m时，评价指标表现不佳，并且从图9可以看出，过小的阈值导致车辆的空间交互性有一定程度的削弱，从而影响了实验效果。当k取35 m时，评价指标略高于25 m时，这表明过大的邻接距离可能引入了冗余和无效车辆信息，干扰了模型的预测，在一定程度上影响了车辆轨迹预测的精度。基于上述分析，k=25 m是一个合适的邻接距离选择。

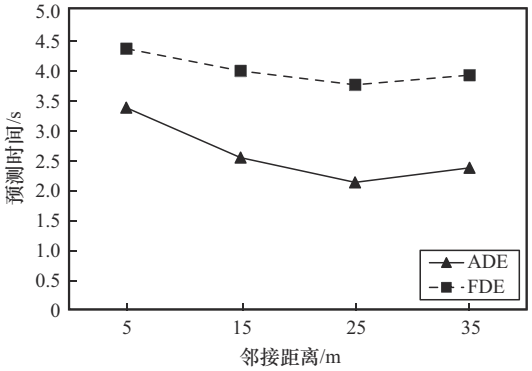
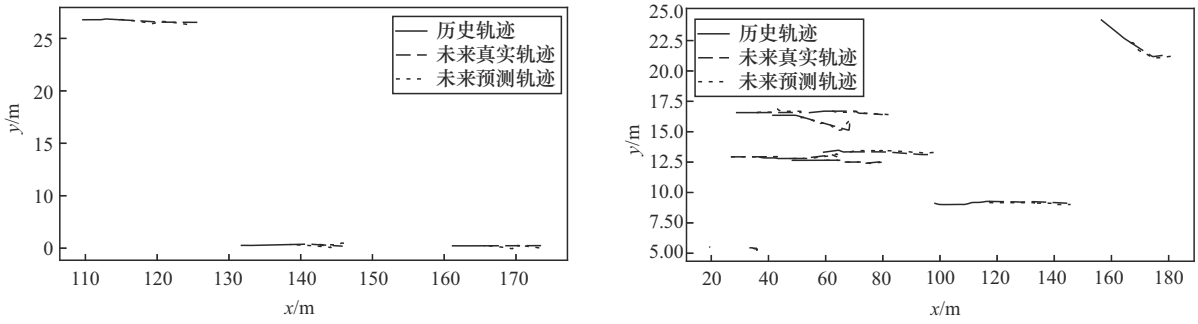


图9 不同邻接阈值k的模型表现比较

3.5 轨迹预测结果可视化

本节对模型预测的未来轨迹进行可视化，以更好地展示实验结果。所有结果数据均取自NGSIM数据集，选取2个场景，分别为车辆稀疏场景和车辆密集场景，结果如图10所示。车辆稀疏场景如图10(a)所示，此场景3辆车的交互较少，基于时空Transformer的网络模型可以更专注于单一车辆的运动状态，展现出较高的预测准确性。图10(b)为车辆密集场景，可以看到，中间两车道密集地行驶着6辆车，其中一辆车行驶过程中企图变道但看到邻接车道有车辆靠近后又迅速取消了变道，这对基



(a) 车辆稀疏场景 (b) 车辆密集场景

图10 轨迹预测可视化

于时空Transformer的网络模型的预测产生了极大的挑战。得益于时空注意力机制，基于时空Transformer的网络模型还是成功预测了取消变道，虽然与真实轨迹有所偏差，但还是能通过车辆间的交互及时做出反应。此外，尽管车辆密集场景下不确定性变得更多，基于时空Transformer的网络模型仍能保持一定的预测精度，适应车辆密集场景的需求。

4　结束语

在深入研究了国内外对轨迹预测的工作现状后，发现目前大部分研究工作都围绕着单个车辆或者局限于单一场景的轨迹预测。为了更准确地实时预测车辆在复杂交通情况下的轨迹，提出了基于时空Transformer的网络模型，该模型的核心优势在于能够捕获所有车辆之间的空间交互作用和时间依赖性。具体方法包括：利用输入车辆的历史特征信息构建时空矩阵，并通过空间自注意力、时间自注意力、时空掩码注意力等在空间层面上对每一个交通参与者之间的潜在关系进行深度挖掘，在时间层面上进行潜在关系表征，并通过特征融合完成对时空特征的提取，最终输出所有车辆的预测轨迹。基于时空Transformer的网络模型在NGSIM数据集上进行了训练和验证，结果表明，与现有的模型相比，该模型在高速公路这样的交通环境下优于现有模型。然而，该模型在其他交通环境，例如十字路口、人行道等的适用性尚未得到验证。未来的工作重点是将该模型进行更广泛的验证。该模型尽管在某些情况下预测性能有所提升，但是否能够完全满足车联网的实时性要求，仍存在一定的不确定性，未来将进一步优化，以确保其在各种交通环境和条件下都能达到所需的实时响应水平。

参考文献：

[1] LUO W J, YANG B, URTASUN R. Fast and furious: real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3569-3577.

[2] 温惠英, 张伟罡, 赵胜. 基于生成对抗网络的车辆换道轨迹预测模型[J]. 华南理工大学学报(自然科学版), 2020, 48(5): 32-40.
WEN H Y, ZHANG W G, ZHAO S. Vehicle lane-change trajectory prediction model based on generative adversarial networks[J]. Journal of South China University of Technology (Natural Science Edition), 2020, 48(5): 32-40.

[3] KHOSROSHAHI A, OHN-BAR E, TRIVEDI M M. Surround vehicles trajectory analysis with recurrent neural networks[C]//Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE Press, 2016: 2267-2272.

[4] PHILLIPS D J, WHEELER T A, KOCHENDERFER M J. Generalizable intention prediction of human drivers at intersections[C]//Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2017: 1665-1670.

[5] KIM B, KANG C M, KIM J, et al. Probabilistic vehicle trajectory prediction over occupancy grid map *via* recurrent neural network[C]//Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE Press, 2017: 399-404.

[6] LEE N, CHOI W, VERNAZA P, et al. DESIRE: distant future prediction in dynamic scenes with interacting agents[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 2165-2174.

[7] 蔡英凤, 朱南楠, 邰康盛, 等. 基于注意力机制的车辆行为预测[J]. 江苏大学学报(自然科学版), 2020, 41(2): 125-130.
CAI Y F, ZHU N N, TAI K S, et al. Vehicle behavior prediction based on attention mechanism[J]. Journal of Jiangsu University (Natural Science Edition), 2020, 41(2): 125-130.

[8] ALAHI A, GOEL K, RAMANATHAN V, et al. Social LSTM: human trajectory prediction in crowded spaces[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 961-971.

[9] GUPTA A, JOHNSON J, LI F F, et al. Social GAN: socially acceptable trajectories with generative adversarial networks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2255-2264.

[10] ZHANG P, OUYANG W L, ZHANG P F, et al. SR-LSTM: state refinement for LSTM towards pedestrian trajectory prediction[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 12077-12086.

[11] DEO N, TRIVEDI M M. Multi-modal trajectory prediction of surrounding vehicles with maneuver based LSTMs[C]//Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2018: 1179-1184.

[12] DEO N, RANGESH A, TRIVEDI M M. How would surround vehicles move? A unified framework for maneuver classification and motion prediction[J]. IEEE Transactions on Intelligent Vehicles, 2018, 3(2): 129-140.

[13] DEO N, TRIVEDI M M. Convolutional social pooling for vehicle trajectory prediction[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2018: 1549-15498.

[14] IVANOVIC B, PAVONE M. The trajectron: probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 2375-2384.

[15] ZHAO Z Y, FANG H W, JIN Z, et al. GISNet: graph-based information sharing network for vehicle trajectory prediction[C]//Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2020: 1-7.

[16] LI X, YING X W, CHUAH M C. GRIP++: enhanced graph-based interaction-aware trajectory prediction for autonomous driving[J]. arXiv Preprint, arXiv: 1907.07792, 2019.

[17] XU D W, SHANG X T, LIU Y, et al. Group vehicle trajectory prediction with global spatio-temporal graph[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(2): 1219-1229.

[18] XU D W, SHANG X T, PENG H, et al. MVHGN: multi-view adaptive hierarchical spatial graph convolution network based trajectory prediction for heterogeneous traffic-agents[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(6): 6217-6226.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceeding of the 31st International Conference on Neural Information Processing Systems. California: MIT Press, 2017: 6000-6010.

[20] GIULIARI F, HASAN I, CRISTANI M, et al. Transformer networks for trajectory forecasting[C]//Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2021: 10335-10342.

[21] YU C J, MA X, REN J W, et al. Spatio-temporal graph transformer networks for pedestrian trajectory prediction[C]//Proceedings of the 16th European Conference on Computer Vision (ECCV 2020). Berlin: Springer, 2020: 507-523.

[22] LIU Y C, ZHANG J H, FANG L J, et al. Multimodal motion prediction with stacked transformers[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 7573-7582.

[23] NGIAM J, CAINE B, VASUDEVAN V, et al. Scene Transformer: a unified architecture for predicting multiple agent trajectories[J]. arXiv Preprint, arXiv: 2106.08417, 2021.

[24] 景荣荣, 吴兰, 张坤鹏. 基于Transformer的自动驾驶交互感知轨迹预测[J]. 科学技术与工程, 2023, 23(26): 11414-11423.
JING R R, WU L, ZHANG K P. Transformer-based interaction-aware trajectory prediction for autonomous driving[J]. Science Technology and Engineering, 2023, 23(26): 11414-11423.

[25] 王庆荣, 谭小泽, 朱昌锋, 等. 基于门控循环单元和Transformer的车辆轨迹预测方法[J]. 汽车技术, 2024(7): 1-8.
WANG Q R, TAN X Z, ZHU C F, et al. Vehicle trajectory prediction method based on GRU and transformer[J]. Automobile Technology, 2024(7): 1-8.

[26] KUEFLER A, MORTON J, WHEELER T, et al. Imitating driver behavior with generative adversarial networks[C]//Proceedings of the 28th IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2017: 204-211.

[作者简介]

赵文红（1981– ），女，河北衡水人，嘉兴南湖学院讲师，主要研究方向为优化计算。

王巍（1980– ），男，博士，河北张家口人，电磁空间安全全国重点实验室研究员，主要研究方向为智能处理、网络优化。

万子璐（2001– ），女，浙江乐清人，浙江工业大学硕士生，主要研究方向为智能驾驶、车辆控制。