基于全局时序注意力的行人轨迹预测方法

成洋 钟珞 *
湖北孝感美珈职业学院，中国·湖北 孝感 432000

摘要：随着自动驾驶、物联网等技术的发展，行人轨迹预测在城市交通疏导、自动驾驶、机器人等领域应用前景广泛。行人运动的突变性与交互复杂性使轨迹预测极具挑战。针对现有 Transformer 模型仅利用前段轨迹特征、忽略后段特征、在不规则运动上预测效果差的问题，本文提出基于全局时序注意力优化的行人轨迹预测方法（GTAO）。GTAO 通过全局时序指导模块实现前、后段序列轨迹的互相优化，并以真实值监督训练，提升鲁棒性与泛化能力。在 ETH 和 UCY 上的实验表明，GTAO 定性与定量结果均优于基线，指标提升 4%~6%。

关键词：行人轨迹预测；Transformer；全局时序注意力；优化

A Human Trajectroy Prediction Method Based on Global Temporal Attention

Cheng Yang, Zhong Luo*
Hubei Xiaogan Meijia Vocational College, China Hubei Xiaogan 432000

Abstract: With the development of autonomous driving and the Internet of Things, pedestrian trajectory prediction has broad applications in traffic management, autonomous driving and robotics. The Transformer models social interactions well, but ignores rear-segment trajectory features, causing poor prediction on irregular motions such as acceleration, deceleration and standing still. To solve this, we propose GTAO: a global temporal attention optimization method with a global temporal guidance module that optimizes the initial prediction using front- and rear-segment features under ground-truth supervision. Experiments on ETH and UCY show GTAO outperforms the baseline by 4% to 6%.

Keywords: Human trajectory prediction; Transformer; Global temporal attention; Optimization

0 引言

随着互联网技术的飞速发展，物联网的发展使我们的生活和工作更加便利与舒适。目前，预测行人未来轨迹有着广泛的应用前景，具体表现为在自动驾驶、视频监控、人群行为模拟、社交机器人、智慧交通等任务上，如图 1、图 2 所示。



图 1 服务机器人应用场景

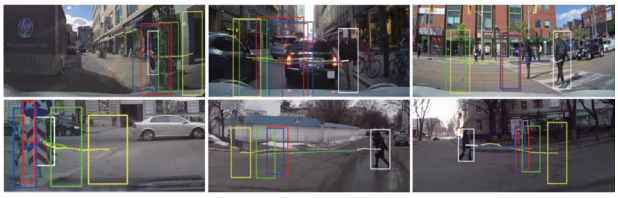


图 2 自动驾驶应用场景

轨迹预测的难点在于行人运动善变：主观上行人往往有明确的目的地，并倾向于选择路程最短的轨迹，其意图难以仅从轨迹数据获取；客观上障碍物、突发事件等环境因素又会使行人突然改变方向、发生方向突变。加之行人运动没有固定的运动模式，行人之间的运动交互复杂多样，容易产生一些难以预测的运动轨迹。

Schneider 等人[1] 用贝叶斯滤波器捕获行人运动状态，提出基于物理运动学的单一模型，但不适用于复杂运动场景。

随着机器学习发展，学者们用行人跟踪算法辅助物理运动学模型，探索复杂离散事件系统状态并应用于轨迹预测。

Pavlovic 等人[2] 提出的切换线性动力学系统（SLDS）能较好建模复杂时空运动，但复杂场景仍需更多数据支撑。

1 基于 Transformer 的行人轨迹预测

Transformer[3] 已经成为自然语言处理（NLP）[4-7] 领域的重要组成部分并取得了显著成功。轨迹坐标点本质上是一段时序数据，这与自然语言处理中的单词序列有着异曲

同工之妙，因此 Transformer 同样适用于轨迹预测，其自注意力机制可以充分挖掘出单个行人历史轨迹特征的时序关系。在时序建模上，前人普遍利用 LSTM 来建模单个行人的历史轨迹特征，而基于 Transformer 的注意力机制能够更好地建模时序轨迹点之间的深层关系。

另有方法在预处理前将轨迹序列构建成图作为 Transformer 输入，使自注意力计算同时考虑时序与空间特征。Yu 等人[8]提出的 STAR 模型完全依赖注意力，融合时序与空间两个 Transformer 模块（如图 3 所示）的特征进行轨迹预测。

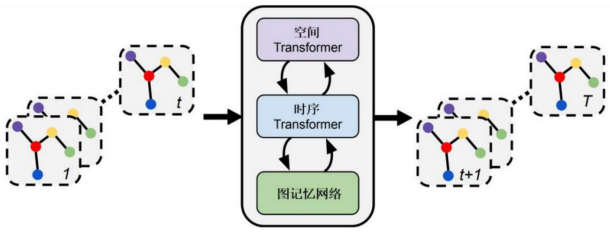


图 3 基于 Transformer 的行人轨迹预测

2 基于 Transformer 的行人轨迹预测任务

基于 Transformer 的行人轨迹预测方法已获广泛关注，本文将在此基础上挖掘轨迹序列的全局注意力特征，降低预测误差。

2.1 问题描述与动机

行人的一段轨迹序列是一段有着很强时序关系的轨迹点，且场景中行人之间的社交交互也影响着行人的后续决策。Transformer 的自注意力机制既能挖掘单个行人的时序历史轨迹特征，也能对目标行人周围其他行人的社交影响程度进行建模，在时空两个层面完成历史轨迹与空间交互特征建模，与行人轨迹预测任务高度匹配。

2.2 基于 Transformer 的行人轨迹预测

2.2.1 Transformer 的基本思想

Transformer 最初由谷歌团队于 2017 年在文献[3]中提出，是一个面向 sequence-to-sequence 任务的模型，最开始主要应用于自然语言处理的各种任务中并取得了很好的效果。模型基于自注意力的原理，可以有效地处理输入与输出，而且比传统的卷积神经网络或递归神经网络更加精确、高效。图 4 为 Transformer 的基本结构，上方为编码器，下方为解码器。

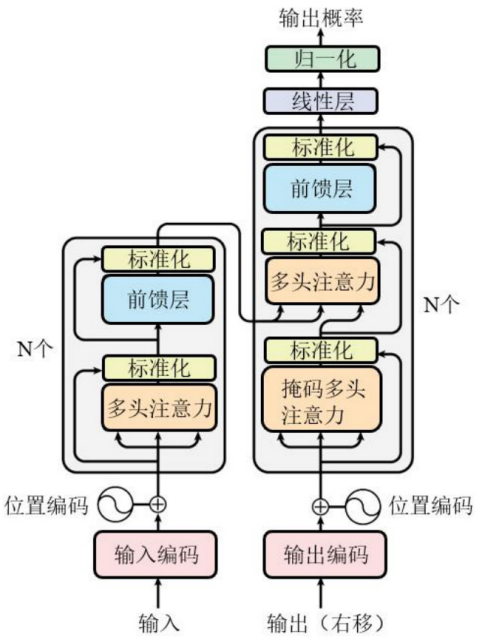


图 4 Transformer 的基本结构

2.2.2 行人轨迹预测基线模型

本文以 Yu 等人[8]在 ECCV 会议上提出的网络框架为基线模型，结构如图 5 所示，也是实验表格中的主要对比对象。

2.3 评价指标和实验设备

2.3.1 行人轨迹预测数据集及评价指标

本文采用的行人轨迹预测数据集包括 ETH[9]和 UCY[10]，其中包含五个子数据集，由定点摄像头拍摄的真实世界行人轨迹所组成，以 2.5fps 的帧率人工标记轨迹二维坐标点，是现有的在行人轨迹预测任务上最常用且公认的公共轨迹数据集，相关信息如表 1 所示。实验采用交叉验证留一法，在四个数据集上训练和验证，并在剩余数据集上进行测试。

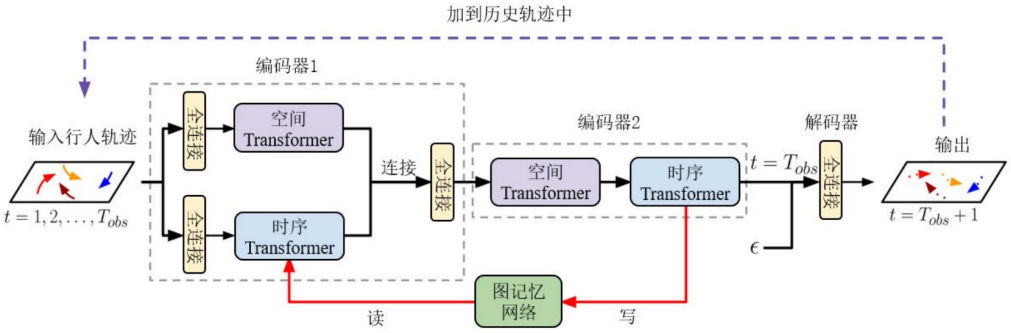


图 5 基线模型的框架结构示意图

表 1 ETH 和 UCY 数据集相关信息

数据集名称	行人数	帧数	环境
ETH	367	12380	室外
HOTEL	420	18060	室外
UNIV	118	7410	室外
ZARA1	148	9010	室外
ZARA2	204	10520	室外

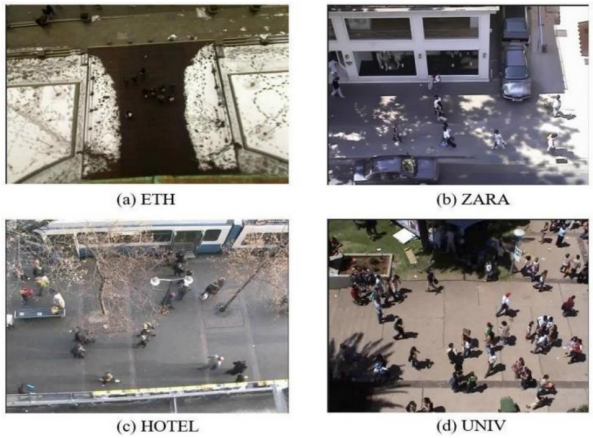


图 6 数据集相关图片示意图

2.3.2 评价指标

为验证模型方法的有效性，本任务采用定量与定性两种评价指标。

本文实验结构评价指标为平均位移误差和最终位移误差。

平均位移误差（Average Displacement Error，ADE），统计场景中所有行人在所有的时间步长内，行人的预测轨迹 $\hat{Y}_n^t$ 与真是轨迹 Y_n^t 之间的 L_2 距离，计算公式如下所示：

$$ADE = \frac{\Sigma_{n\in N}\Sigma_{t\in TP}\left\|\hat{Y}_n^t - Y_n^t\right\|_2}{N * T_P} \quad (1)$$

最终位移误差（Final Displacement Error，FDE），统计场景中所有行人最后 一个时间步的轨迹坐标与真实值之间的 L_2 距离，计算公式如下所示：

$$FDE = \frac{\Sigma_{n\in N}\left\|\hat{Y}_n^t - Y_n^t\right\|_2}{N},\ t = T_p \quad (2)$$

2.3.3 实验环境

在本文的实验环境中，具体的设备参数信息如表 2 所示：

表 2 硬件设备以及相关配置

项目	配置
操作系统	CentOS Linux release 7.7.1908
内核	3.10.0-862.el7.x86_64
CPU	28 个物理核，56 个逻辑处理器，
	主频：800.048MHz，Intel(R) Xeon(R) Gold 5117 CPU
内存	220GB, 2400MHz
GPU	2 块 NVIDIA Tesla PH402 SKU 200

3 基于全局时序注意力优化的行人轨迹预测

本章利用全局时序注意力方法提取整段轨迹序列特征，在保留初始预测信息的同时进行特征优化与轨迹矫正，提升建模全局特征的能力。

3.1 问题描述与动机

现有基于 Transformer 的方法虽能建模社交行为，但只考虑前段轨迹特征，忽略了整段序列前后特征的强关联性，未能充分利用全局时序轨迹特征。

如图 7 所示，这会导致模型在加速、减速、静止等一些不规则的运动上带来一定程度的性能降低。本章提出的方法能够充分利用后续轨迹序列特征，利用全局注意力机制对初始预测进行性能优化，有效提高模型的泛化能力以及异常处理的能力。

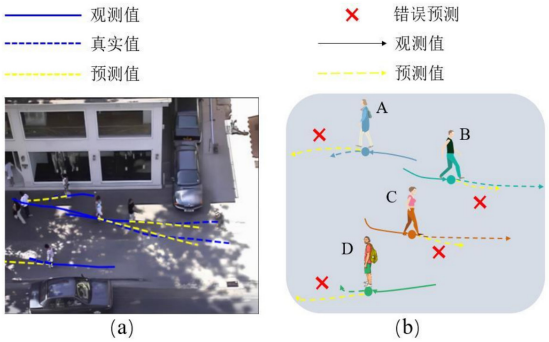


图 7 基于真实场景的行人轨迹预测分析示意图

3.2 基于全局时序注意力优化的行人轨迹预测

为解决上述问题，本章提出基于全局时序注意力优化的行人轨迹预测网络（GTAO）。

如图 8 所示，整体框架分为三个部分：基线模型的时

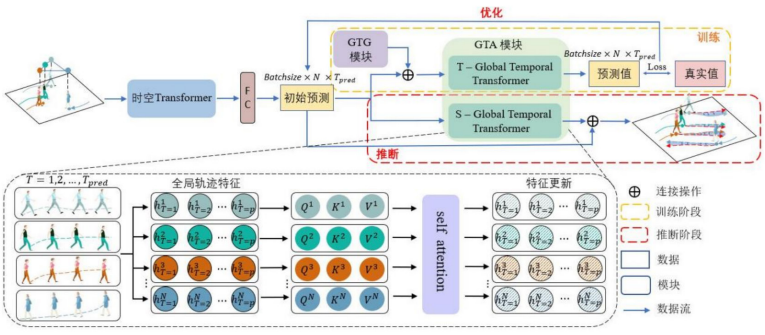


图 8 基于全局时序注意力优化的行人轨迹预测框架 GTAO 示意图

空 Transformer 模块、全局时序注意力模块（GTA Module）与全局时序指导模块（GTG Module）。时空 Transformer 的输入是一段轨迹点坐标序列，输出是初步的整体预测轨迹；在训练阶段，初步的预测被送入 GTA 模块，在 GTG 模块的指导下对时空 Transformer 的初始预测进行优化；在推理阶段，未来的预测轨迹由 S-GTA 得到。

经过全局时序注意力模块得到的优化结果 tra_i^T 和 tra_i^s 之后，为了在训练阶段能够指导模型在基线模型的基础上来进行优化，以此来让模型来学习更加接近真实轨迹的特征，能够更好的解决在某些异常运动中的问题，介绍了一个全局轨迹序列损失以此来训练整个模型，可以被定义为：

$$\begin{cases} L_{gts}^{T} = \sum_{i=1}^{N} \left\| tra_i^T - Y_i \right\|_2 \\ L_{gts}^{S} = \sum_{i=1}^{N} \left\| tra_i^S - Y_i \right\|_2 \end{cases} \tag{6}$$

其中，L_{gts}^{T} 和 L_{gts}^{s} 代表优化结果 $tra_i^T | tra_i^s$ 和真实轨迹 Y_i 之间的平均均方误差。通常情况下，基线模型的训练过程也有一个损失，被定义为：

$$L_{nor} = \sum_{i=1}^{N} \left\| \widehat{tra}_i - Y_i \right\|_2 \tag{7}$$

$\widehat{tra}_i$ 是 baseline 的初始化预测结果，为了确保上述损失的共同有效性，本章最终 使用结合损失如下：

$$L = L_{nor} + L_{gts}^{T} + L_{gts}^{S} \tag{8}$$

由于全局时序注意力模块主要是提取全局轨迹注意力特征，这就可能丢失了基线模型初始预测的局部轨迹重要信息，因此需要对基线模型的初始预测和经过全局时序注意力模块的优化结果来进行融合，相关过程定义为：

$$F_i = \lambda_1 \widehat{tra}_i + \lambda_2 tra_i^S \tag{9}$$

其中，λ_1 和 λ_2 是融合基线模型和 S-GTA 预测结果的权重大小，相关消融实验结果将在后续小节中呈现。

3.2.1 全局时序指导模块

全局时序注意力模块在进行优化时可以处理整个轨迹序列，提取整体轨迹点序列之间的相关性，但是仅使用 GTA 模块对模型的最终结果是没有贡献的。如果只利用基线模型的初始预测作为 T-GTA 的输入，将不会得到优化的结果，因为基线模型的初始预测在某些步骤中可能是不正确的，如果不加以更正，则模型会在后续建模中的偏离误差会越来越大。因此，在训练阶段中，引入了一个全局时序指导模块来与 T-GTA 协作，整体模型在 GTG 的指导帮助下，来学习更接近真实轨迹的预测结果 embedding。如图 3-4 所示，从基线模型获得的初始预测结果 $\widehat{tra}_i$，将其与真实的轨迹映射到相同的特征空间中，将嵌入的特征连接起来，得到融合的特征嵌入的特征表达，可以表述为：

$$tra_i = \sigma\left(\widehat{tra}_i \oplus tra_{gt}\right) \tag{10}$$

其中表示真实轨迹 embedding，σ 表示 sigmod 函数，表示包含真实轨迹 指导效果的 embedding。轨迹特征 embedding 是由训练阶段的基线模型初始预测结果和地面真实轨迹相结合所由此组成的，由此可以指导模型处理一些不规则的运动，训练过程中通过反向传播机制对 GTA 和 GTG 来进行监督训练。

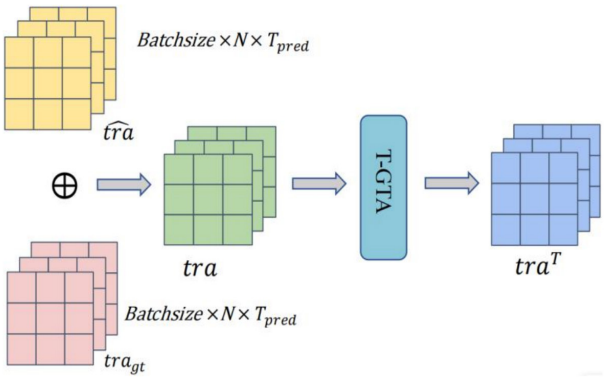


图 9 GTG 模块的初始预测和真实轨迹的特征融合示意图

3.3 实验参数与分析

3.3.1 实验参数

GTAO 用多层感知机（MLP）将轨迹坐标嵌入到 32 维，经 ReLU 校正；批量规模 16，Adam 学习率 0.0015，丢失率 0.1，训练 300 个 epoch，时序与空间 Transformer 编码器层数分别为 1 和 2，全局时序 Transformer 编码器 2 层、每层 8 头、512 维；融合权重 λ_1、λ_2 均设为 0.5。

3.3.2 预测结果

行人轨迹预测任务是根据输入的前 8 帧轨迹点坐标来预测后 12 帧轨迹点坐标。为了与基线模型和现有方法进行一致性比较，在两个标准且公认的行人轨迹数据集 ETH 和 UCY 上，分别在 ADE 和 FDE 两个评价指标上保留三位有效数字，计算结果为误差大小，数据越小代表指标越好。其中确定性结果（Deterministic）表示对于每个行人采样 1 次，模型所预测的结果。

随机性结果（Stochastic）指对每个行人采样 20 次，取指标最好的结果。本文在两种层面上与基线及其他方法比较，无结果以“-”表示，对比方法如表 3 所示。

最好的结果由粗体表示。

如表 3 所示，与 9 种先进方法相比，本文模型在平均 AVG 误差的 ADE 指标上取得最好结果，在 ETH 和 HOTEL 上 ADE 均最佳，其余数据集上亦有竞争力。

在 UNIV 数据集上本文结果与 CIDNN 相差较大，

表 3 对比方法

方法	会议 / 期刊	ETH	HOTLE	UNIV	ZARA	ZARA2	AVG
Social-LSTM-D [32]	CVPR 16	0.77/1.60	0.38/0.80	0.58/1.28	0.51/1.19	0.39/0.89	0.53/1.15
SocialAttention [60]	ICRA 18	1.39/2.39	2.51/2.91	0.88/1.75	1.25/2.54	1.01/2.17	1.41/2.35
CIDNN [61]	CVPR 18	1.25/2.32	1.31/1.86	0.51/1.07	0.90/1.28	0.50/1.04	0.89/1.73
Social-GAN(1V-1) [37]	CVPR 18	1.13/2.21	1.01/2.18	0.60/1.28	0.42/0.91	0.52/1.11	0.74/1.54
TrafficPredict [62]	AAAI 19	5.46/9.73	2.55/3.57	3.31/6.37	4.32/8.00	3.76/7.20	3.88/6.97
Social-BIGAT(K=1) [39]	NeurIPS 19	–	–	–	–	–	0.61/1.33
STGAT(1V-1) [41]	ICCV/19	0.88/1.66	0.56/1.15	0.52/1.13	0.41/0.91	0.31/0.68	0.54/1.11
PITF(Single-model) [63]	CVPR 19	0.88/1.98	0.36/0.74	0.62/1.32	0.42/0.90	0.34/0.75	0.52/1.14
STAR-D(基线模型)[51]	ECCV 20	0.66/1.33	0.44/0.97	0.60/1.33	0.50/1.12	0.39/0.90	0.52/1.13
GTAO-D(本文方法)		0.64/1.37	0.36/0.81	0.63/1.39	0.49/1.10	0.41/0.93	0.51/1.12

表 4 ETH 和 UCY 上的随机性实验结果

方法	会议 / 期刊	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Social-LSTM [32]	CVPR 16	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
Social-GAN [37]	CVPR 18	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
Sophie [64]	CVPR 19	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.54/1.15
STGAT [41]	ICCV 19	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
TPNMS [65]	AAAI 21	0.52/0.89	0.22/0.39	0.55/1.13	0.35/0.70	0.27/0.65	0.38/0.73
STAR(基线模型)[51]	ECCV 20	0.36/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53
GTAO (本文方法)		0.34/0.63	0.17/0.35	0.29/0.60	0.23/0.48	0.20/0.42	0.25/0.50

表 5 消融实验——ETH 和 UCY 上各模块的消融实验对比

基线模型	全局时序注意力模块	全局时序指导模块	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
√	×	×	0.36/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53
√	√	×	0.38/0.68	0.20/0.40	0.35/0.70	0.26/0.57	0.24/0.50	0.28/0.57
√	√	√	0.34/0.63	0.17/0.35	0.29/0.60	0.23/0.48	0.20/0.42	0.25/0.50

表 6 消融实验——ETH 和 UCY 上 1 和 2 权重的消融实验对比

λ_1	λ_2	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
0.7	0.3	0.35/0.65	0.17/0.35	0.30/0.63	0.24/0.53	0.20/0.42	0.25/0.52
0.3	0.7	0.35/0.65	0.18/0.40	0.29/0.60	0.26/0.54	0.21/0.42	0.26/0.52
0.5	0.5	0.34/0.63	0.17/0.35	0.29/0.60	0.23/0.48	0.20/0.42	0.25/0.50

因其行人密度大，需模型具备处理高密度人群的能力；CIDNN 擅长高密度建模，在稀疏场景中效果较差，而本文方法更适用于稀疏场景，故平均误差 AVG 明显更优。

与基线相比，本文方法确定性结果的 AVG 指标提升 2.0% 和 0.9%；如表 4 所示，随机性层面在五个子数据集上均取得最好结果，AVG 较基线提升 4.0% 和 6.0%，表明全局时序注意力对理解行人全局运动轨迹至关重要。

消融实验结果如表 5 所示，√表示纳入该模块。第一行显示仅用基线模型时效果较差；对比前两行，单独加入全局时序注意力模块后效果反而变差，因为缺少全局时序指导模块时，基线初始预测的错误经全局时序注意力建模后会越偏越大。

由最后一行可知，全局时序指导模块与全局时序注意力模块协作时，模型能充分提取全局轨迹序列点间的注意力与特征关系，有效解决异常运动，取得更好的指标，说明两模块共同参与能降低预测误差、提升预测性能。

全体模块加入后，模型输出考虑了全局轨迹特征，但可能丢失基线预测中的细粒度信息，故本章对初始预测与全局优化预测进行融合参数消融实验。如表 6 所示，当 λ_1、λ_2 均为 0.5 时平均 ADE/FDE 达最优 0.25/0.50；初始预测权重更大（0.25/0.52）或全局优化权重更大（0.26/0.52）时结果均较差。

4 结语

近年来，物联网从耳机、手环到家居机器人、自动驾驶、无人机，使生活工作更加便捷。对自动驾驶而言，准确预测行人运动轨迹以防止碰撞是重要环节，在自动驾驶普及前亟需解决。

现有基于 Transformer 的方法虽能建模社交行为，但只利用局部注意力，忽略前后关联的全局时序轨迹特征，导致在不规则运动上性能下降。为此，本文提出 GTAO：以全局时序注意力优化模块扩大感受野，充分利用整段序列的前后状态优化预测；并引入全局时序指导模块监督训练，避免基线初始预测误差在 GTA 作用下不断放大。

在 ETH 和 UCY 两个公认的标准数据集上，经过实

验结果、模块消融结果和清晰直白的可视化预测结果等分析，验证了本文提出的 GTAO 模型的准确性。在最终的平均 ADE 和 FDE 指标上，本文方法相较于基线模型分别在确定性结果和随机性结果上提升了 2.0% 和 0.9% 以及 4.0% 和 6.0%。与基线模型相比，GTAO 模型的预测结果更加接近于真实轨迹。

参考文献：

[1] Schneider N, Gavrila D M. Pedestrian path prediction with recursive bayesian filters: A comparative study[C]. Pattern Recognition: 35th German Conference, GCPR September 3-6, 2013: 174-183.

[2] Pavlovic V, Rehg J M, MacCormick J. Learning switching linear models of human motion[J]. Advances in neural information processing systems, 2000, 13.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[4] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[5] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.

[6] Yang Z, Dai Z, Yang Y, et al. generalized autoregressive pretraining for language understanding; 2019[J]. Preprint at https: arxiv. org abs 1906.08237 Accessed June, 2021, 21.

[7] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing[J]. IEEE Computational intelligenCe magazine, 2018, 13(3): 55-75.

[8] Yu C, Ma X, Ren J, et al. Spatio-temporal graph transformer networks for pedestrian trajectory prediction[C]. Proceedings of the IEEE CVF European Conference on Computer Vision(ECCV), 2020: 507-523.

[9] Pellegrini S, Ess A, Schindler K, et al. You'll never walk alone: Modeling social behavior for multi-target tracking[C]. Proceedings of the IEEE CVF International Conference on Computer Vision(ICCV). 2009: 261-268.

[10] Lerner A, Chrysanthou Y, Lischinski D. Crowds by example[C]. Computer graphics forum. Oxford, UK: Blackwell Publishing Ltd, 2007, 26(3): 655-664.

作者简介：成洋（1981.12-），男，汉族，湖北武汉人，本科学历，湖北孝感美珈职业学院，讲师，研究方向：网络安全、智能方法与系统。

* 通讯作者：钟珞（1957.06-），男，汉族，湖北武汉，硕士学历，武汉理工大学，教授，研究方向：智能方法与系统、知识发现与数据挖掘、软件工程。