VIT-Ped：用于行人行为分析的视觉意图变换器

Aly R . Elkammar[1,2,+]，**Karim M. Gamaleldin**[1,2,+]，以及 **Catherine M. Elias**[1,2,*]

[1]德国开罗大学媒体工程与技术学院，计算机科学与工程系，埃及开罗[2]C-DRiVeS实验室：车辆系统认知驾驶研究，埃及开罗[*]catherine.elias@ieee.org[+]这些作者对本研究的贡献相同

arXiv:2601.01989v1 [cs.CV] 5 Jan 2026

摘要

行人意图预测是从3级向4级自动驾驶过渡过程中的关键技术之一。为了理解行人过街行为，需要考虑多种因素和特征，从而让未来的道路对每个人都更加安全。我们提出了一种基于 Transformer/视频视觉 Transformer 的算法，并设计了不同规模的模型，以使用不同的数据模态。我们在广受欢迎的行人行为数据集 JAAD 上评估了这些算法，取得了最先进性能，并在准确率、曲线下面积和 F1分数等评估指标上达到最先进水平。我们通过广泛的消融研究，考察了不同模型设计选择带来的优势。

引言与相关工作

过去几年中发生了许多交通事故，其中 94% 是由人为错误造成的。例如，25% 的致命事故是由超速造成的。这些问题表明，我们需要实现驾驶活动的自动化。因此，预计到 2045 年，美国 60% 的汽车将实现完全自动驾驶。然而，提升汽车自动化水平也面临诸多挑战。根据世界卫生组织（WHO）的数据，每年约有 135 万人死亡，另有 2,000 万至 5,000 万人受伤。若缺乏恰当的安全措施和算法，这些数字可能会急剧攀升，使城市环境变得前所未有地危险。因此，针对不同城市场景开发算法是研发工作的重点。除了障碍物检测和高级驾驶辅助系统（ADAS）之外，行人意图／行为预测也至关重要，因为它能预判行人移动，从而帮助及时采取更恰当的决策，让未来的道路对每个人都更加安全。

过去十年来，行人意图预测一直是一个热门研究课题，人们尝试使用不同的数据集、算法和特征来应对这一挑战。然而，由于人类行为极其复杂，意图也难以预测，这项任务至今仍受到广泛研究，人们不断努力争取尽可能好的性能。在文献中，意图、动作和行为这几个术语常被互换使用，用来描述行人行为。在我们的工作中，我们试图预测行人的意图；意图可定义为代表行人承诺与计划的心理状态，而经典方法无法对其进行观测。

多年来，人们提出了许多算法、模态和特征来解决这项任务。在早期，由于图形处理器技术的局限，支持向量机（SVM）是这类任务的首选模型。例如[1] 采用小型卷积神经网络对视觉数据进行编码，再将其输入支持向量机模型，以分类过街和不采取过街行为。近期的研究则使用输入序列，以便同时捕捉空间和时间维度，并利用序列中不同帧之间的依赖关系。这种方法称为**时空建模**，可通过多个步骤实现。首先，我们需要使用前馈神经网络（全连接层）提取或编码非视觉数据的空间特征，并使用二维卷积神经网络（CNN）或图卷积神经网络处理视觉数据。随后，需要将这些特征输入能够捕捉时间依赖关系的架构，例如循环神经网络（RNN）、长短期记忆网络（LSTM）和门控循环单元（GRU）。例如，[2] 使用二维卷积神经网络对视觉特征进行编码，然后将编码后的特征传入循环神经网络，以编码不同帧之间的时间特征，最后将编码后的时空特征输入全连接层进行动作分类。

与上述方法不同，文献中还有其他方法尝试同时学习时空特征。例如[3] 和[4] 使用了 3D 卷积神经网络，它能够捕捉时空特征，而其二维对应模型则无法做到这一点。为此，它将负责卷积和池化运算的二维卷积核替换为 3D 版本

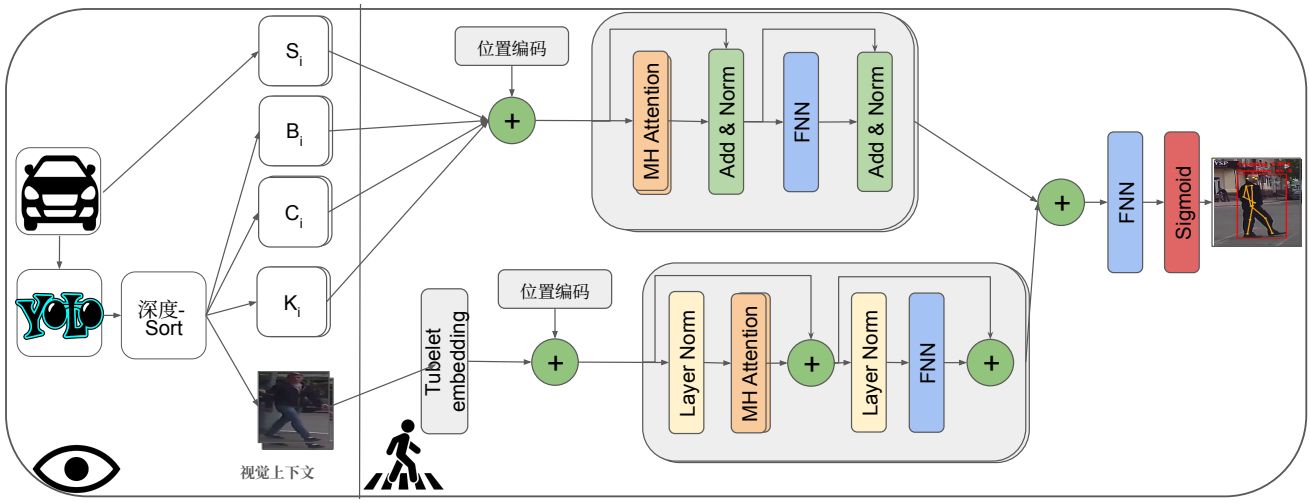

图 1. VIT-Ped 架构概览

。[5] 采用的另一种方法是扩展长短期记忆网络，加入一种称为 ConvLSTM 的卷积结构，从而有效学习空间和时间特征。

最近，2017年 Transformer 被引入自然语言翻译领域，这标志着深度学习领域的一项重大发展。它们不仅在自然语言处理领域超越了其他架构，而且经过若干改进后[6] 也主导了计算机视觉领域。这种主导地位主要归功于注意力机制的应用，该机制能够衡量输入数据不同部分的重要性，从而实现并行处理，并捕获序列中的长程依赖，且不会造成信息损失。近年来，文献中出现了基于 Transformer 的算法，其中[9] ,[10] 和[11] 采用这类算法进行行人意图预测，无需再使用循环神经网络和/或卷积神经网络。另有研究者[3] 和[12], 尝试将注意力机制与常规循环神经网络和卷积神经网络结合，以解决当前任务。

在这项工作中，我们提出了一种新方法 PedViViT。该方法对最先进的视频视觉 Transformer 模型[8] 进行了改进，用于处理视觉数据，并采用基于 Transformer 的编码器[6] 处理非视觉数据，从而预测行人是否会过马路。总而言之，本文的贡献如下：

1. 一种用于行人意图预测的多模态模型。该模型基于一种名为视频视觉 Transformer（ViViT）的 Transformer架构模型[8]。
2. 使用小型模型实现最先进性能（比 PCPA 小 6 倍），更适用于实时应用。
3. 使用非视觉模型实现最先进性能（比 PCPA 小 15 倍），该模型仅使用边界框坐标等数值输入。
4. 消融研究探索了不同的特征和模态。

方法

问题定义

行人意图预测任务定义如下：给定最近 n 帧的数据序列，我们需要预测和估计行人穿越道路的可能性，以帮助自车做出更好的决策。我们的目标是使用过去 n 个时间戳来预测穿越的概率 $C_i \in \{0,1\}$。

在不同的模型中，我们使用不同的特征来帮助模型为每个行人预测C_i。模型使用的特征包括：

- **边界框**：表示行人 i 的左上角和右下角坐标：

$$B_i = \{b_i^{t-m}, b_i^{t-m+1}, ..., b_i^t\}$$

- **中心**：表示行人 i 的边界框中心：

$$C_i = \{c_i^{t-m}, c_i^{t-m+1}, ..., c_i^t\}$$

- **姿态关键点**：表示行人 i 的 18 个关键点

$$K_i = \{k_i^{t-m}, k_i^{t-m+1}, ..., k_i^t\}$$

- **自车速度**：

$$S_i = \{s^{t-m}, s^{t-m+1}, ..., s^t\}$$

- **局部上下文**：显示行人 i 周围情况的帧裁剪版本

$$L_i = \{l_i^{t-m}, l_i^{t-m+1}, ..., l_i^t\}$$

- **全局上下文**：完整帧

$$G_i = \{g^{t-m}, g^{t-m+1}, ..., g^t\}$$

- **局部周边**：对行人 i 进行遮罩后的帧裁剪版本，使其像素全部呈现为同一种颜色

$$A_i = \{a_i^{t-m}, a_i^{t-m+1}, ..., a_i^t\}$$

每种输入类型都有一个长度为 m+1 的序列，而 t 是穿越或不穿越行为发生前 1—2 秒范围内的任意时间点。

数据准备与预处理

在我们的实验中，我们使用自动驾驶中的联合注意力（JAAD）[1] 数据集来获取上述特征。JAAD 数据集包含 346 段短视频片段（时长 5–10 秒），这些片段从在北美和东欧拍摄的 240 多小时驾驶录像中提取而来。视频由车载摄像机以 30 帧率录制，记录了 1k 名不同的行人。

不同的输入类型需要采用不同的预处理方法。

边界框与中心

这些特征能告诉我们行人当前在帧中的位置，并提供更多关于其在连续帧间移动情况的信息。我们采用一种称为增量编码的技术对这些特征进行归一化：将第一帧中的边界框和中心坐标从后续 n-1 帧的边界框和中心坐标中减去，然后移除第一帧中的边界框 B 和中心 C。每个边界框都是一个由二维坐标组成的向量，其中包含左上角和右下角的坐标，即：

$$b_i^{t-m} = \{x_{itl}^{t-m}, y_{itl}^{t-m}, x_{ibr}^{t-m}, y_{ibr}^{t-m}\}$$

其中 x_{itl}，y_{itl}表示左上角点，而 x_{ibr}，y_{ibr}表示右下角点。

每个中心都是由单个二维坐标构成的向量，即：

$$c_i^{t-m} = \{x_i^{t-m}, y_i^{t-m}\}$$

姿态关键点

姿态关键点表示目标行人在每一帧中的姿态和详细运动情况。然而，JAAD 数据集未提供真实姿态关键点，而是由[3]中提到的PedestrianActionBenchmark仓库提供。单个姿态关键点向量 k 是一个由二维坐标构成的36维向量，其中包含18个姿态关节，即：

$$k_i^{t-m} = \{x_{i1}^{t-m}, y_{i1}^{t-m}, ..., x_{i18}^{t-m}, y_{i18}^{t-m}\}$$

自车速度

JAAD 提供自车速度 S；不过，该速度被分为5类：停止、缓慢行驶_、快速行驶_、减速和加速。

局部上下文

局部上下文 Lij 提供与目标行人相关的视觉特征，以及其周围的一些环境信息，前提是行人 i 的边界框放大比例大于 1。Lij 由一系列大小不同的 RGB 图像组成，这些图像覆盖目标行人周围（[112,112],[128,128],[224,224]）像素的区域。图像通过除以 255 进行归一化。

局部周边

局部周边 A^{ij} 与局部上下文相似，但不同之处在于，目标行人会被遮蔽，即将该行人对应的边界框灰显，而周围环境保持不变。图像通过除以 255 进行归一化。

全局上下文

全局上下文 G^{ij} 表示整个画面，该画面被调整为以下尺寸之一（[112,112],[224,224]）像素。图像通过除以 255 进行归一化。此输入的目的是提供所需的视觉特征，以展现不同道路使用者之间以及他们与道路之间的各种互动。

感知

在深入介绍我们的模型架构之前，我们先搭建了测试环境。我们使用了不同场景的预录视频，其中行人会穿过或不穿过道路，以检验我们的模型能否正确预测其意图，以及能否提前 >=1 秒作出预测。部分视频来自 JAAD 数据集，我们也在 GUC 校园自行录制了一些视频。因此，我们使用 Ultralytics 的 YOLOv8 来检测行人、估计其姿态关键点，并为每位行人分配唯一 ID 以进行跟踪。随后，我们准备好待测模型的输入并将其送入模型，测试模型能否正确预测行人的意图。

模型架构

我们的模型使用两种数据模态。对于车辆速度、边界框、中心和关键点坐标等非视觉数据，我们使用 Transformer[6]。对于局部上下文、全局上下文和局部周边等视觉数据，我们使用视频视觉 Transformer（ViViT）[8] 来处理这类数据。上述不同 Transformer 得到的特征会以不同方式融合。我们使用的模块包括：

Transformer 模块

我们对所有非视觉数据使用原始 Transformer 编码器[6] 。首先，我们沿非视觉数据的最后一个维度进行拼接，然后对其应用位置编码（2）（所用的位置编码来自[6]），以保持非视觉数据在连续帧间的序列顺序，使输入成为形状为（批量大小、帧数、特征）的三维数组。随后，我们将编码后的数据输入包含 N 个编码器层的 Transformer 编码器，使自注意力机制能够对输入序列不同部分的重要性进行加权，并捕捉它们之间的不同关系。

$$P = \begin{cases} \sin\left(\frac{\text{pos}}{10000^{2i/d}}\right), & \text{if } i \mod 2 = 0 \\ \cos\left(\frac{\text{pos}}{10000^{2i/d}}\right), & \text{if } i \mod 2 = 1 \end{cases} \tag{1}$$

ViViT 模块

对于视觉输入，我们首先对输入数据执行一种称为 Tubelet 嵌入[8] 的特殊嵌入操作，以捕捉视频样本中某个图块随时间变化的信息，这是对[7]中基于图块的嵌入的扩展。输入数据是由一系列 RGB 图像组成的序列，以形状为（批量大小、帧数、图像高度、图像宽度、通道数）的五维数组形式收集。随后，将位置编码应用于这些嵌入，结果被输入到 ViViT[8]模型中，其中每种视觉输入类型（局部周边、局部上下文和全局上下文）都具有独立的 ViViT[8] 模型。

融合模块

将数据输入上述模块后，需要以某种方式将它们的输出融合起来，才能输出行人即将横穿马路的意图概率。我们探索了多种数据融合方式，其中一些方法是沿特征维度拼接输出，然后将结果输入前馈神经网络、全局平均池化层、另一个原始 Transformer 编码器或长短期记忆网络，以学习有关输入数据的更多时间特征。最后，我们将融合模块输出的结果输入 sigmoid（2）层，以输出行人是否即将横穿马路的意图概率。

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

其中 x 表示融合模块的输出。

模型	JAAD 全部					JAAD 行为				
	Acc	AUC	F1	精确率	召回率	Acc	AUC	F1	精确率	召回率
ConvLSTM	0.59	0.58	0.33	0.23	0.58	0.59	0.53	0.70	0.64	0.77
CAPformer	0.73	0.78	0.53	0.38	**0.85**	0.51	0.54	0.52	**0.68**	0.42
PCPA	0.76	0.78	0.55	0.41	0.82	0.54	0.49	0.65	0.62	0.68
C3D	0.85	0.76	0.60	0.57	0.63	0.58	0.48	0.73	0.61	0.91
我们的方法 1	0.81	0.74	0.51	0.47	0.55	0.62	**0.61**	**0.77**	0.63	**0.99**
我们的方法 2（非视觉）	0.84	**0.81**	0.51	0.52	0.50	**0.64**	0.58	**0.77**	0.65	0.93
我们的方法 3	**0.86**	0.77	**0.61**	**0.58**	0.64	0.62	0.55	0.73	0.66	0.83

表1. JAAD 数据集评估指标中各类模型的比较

模型	JAAD 全部					JAAD 行为				
	Acc	AUC	F1	精确率	召回率	Acc	AUC	F1	精确率	召回率
Ours 4（因式分解编码器）	0.82	0.79	0.59	0.49	0.75	0.60	0.53	0.71	0.64	0.81
Ours 5（预训练）	0.83	0.85	0.56	0.50	0.63	0.61	0.64	0.67	0.61	0.63
Ours 6（仅边界框）	0.78	0.73	0.51	0.42	0.65	0.63	0.56	0.74	0.66	0.83
Ours 7（集成）	0.75	0.79	0.54	0.40	0.84	0.64	0.58	0.74	0.68	0.82
我们的方法 8（微调与非视觉）	0.75	0.70	0.46	0.37	0.62	0.53	0.54	0.58	0.66	0.52
我们的方法 9（注意力掩码）	0.81	0.74	0.51	0.47	0.55	0.62	0.61	0.77	0.63	0.99
我们的方法 10	0.82	0.79	0.59	0.48	0.75	0.51	0.51	0.57	0.64	0.52
我们的方法 11	0.84	0.78	0.60	0.53	0.70	0.56	0.57	0.59	0.70	0.52
我们的方法 12	0.82	0.79	0.59	0.49	0.74	0.52	0.48	0.62	0.61	0.63

表2. 消融研究

实验

数据集

我们的模型在 JAAD_{all} 和 JAAD_{beh} 上进行了训练和评估，这两个数据集是 JAAD 数据集的两个子集[1]。它由简短的驾驶自车视频片段组成，其中包含边界框和车辆标注等标注信息。在我们的工作中，我们添加了行人关键点，用于捕捉行人身体上 18 个关键点的位置。我们的模型经过训练，能够准确预测行人的过街意图，预测时间窗口为实际过街动作发生前 1 至 2 秒。使用了[13] 中介绍的训练集。

训练

我们的小型模型使用初始学习率 $3e^{-4}$ 进行训练，批量大小为 32，使用 Adam优化器训练 100 个训练轮次。耐心值设为 5 个训练轮次；如果学习率趋于平稳，则将学习率降低至原来的 0.2 倍。此外，耐心值设为 15 个训练轮次时，我们使用早停回调，在模型开始过拟合时立即停止训练。为应对 JAAD 数据集中的数据不平衡问题，我们在损失函数（二元交叉熵函数（3））中使用了类别权重。我们的较大模型（表1中的 Ours 3）使用两个数据集进行训练，初始学习率为 $5e^{-5}$，批量大小为 8，使用 Adam优化器训练 20 个训练轮次；我们还使用类别权重来减轻数据集不平衡的影响。随后，我们对在 JAAD_{beh} 上训练的模型进行了 20 个训练轮次的微调，但未使用类别权重；不过，我们使用了 ReduceLROnPlateau回调，耐心值为 5 个训练轮次。如果验证损失未改善，之后便将学习率降低至原来的 0.1 倍。

$$L(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{3}$$

评估指标

对于我们的任务（行人意图预测），我们使用常见的分类指标，例如：准确率、曲线下面积（AUC）、F1分数、精确率和召回率。

模型
使用[3] 提出的代码库，我们将自己的模型与行人意图预测领域的最先进模型进行比较。对比的模型包括：ConvLSTM[5], CAPformer[9],PCPA[3] 和 C3D[4]。

消融研究
如表 2所示，我们开展了消融研究，以比较不同的融合策略，并尝试不同的 ViViT[8] 架构，包括时空注意力和因式分解编码器模型。此外，我们还尝试了不同的图像尺寸以及不同类型输入的组合，以观察模型如何处理融合在一起的不同模态。我们也尝试了 FT-Transformer 中针对表格数据（本文将其视为非视觉数据）的特征标记器方法[13] ，并在对 JAAD 数据集中的样本进行重采样、以平衡数量较少的类别中的正负样本后，使用该数据集训练了一些模型。以下是我们实验中最重要的模型：

- **我们的模型 4**：我们对局部周边和全局上下文输入采用因式分解编码器架构，对其余非视觉数据（边界框、姿态、中心）采用标准 Transformer，并通过全局平均池化将输出融合在一起。

- **我们的模型 5**：我们没有仅使用事件发生前 30–60 帧训练模型，而是使用整个数据集对模型进行预训练，使其能够从事件发生前最早 500 帧开始预测通过行为。随后，我们尝试了多种微调方法，以提升模型在该数据集上的准确率和效率。

- **Ours 6**：我们训练了一个包含 2 层和 4 个头的原始 Transformer 编码器，仅使用边界框作为输入，正如 ts.
- **Ours 7**：该模型是由另外 3 个模型组成的集成。我们先自行对这些模型进行预训练，然后冻结这些模型，并添加一个 Sigmoid 层。该层接收这些冻结的预训练模型的输出，并经过训练以输出正确的概率。其中两个模型分别在两个数据集的平衡版本上进行训练，其中一个是因式分解编码器模型（Ours 10），另一个是时空注意力模型（Ours 12）。然而，第三个模型是在数据集的原始版本上训练的（Ours 11）。
- **Ours 8**：我们使用 FT-Transformer 处理非视觉数据（边界框、姿态、中心、速度），将数值特征转换为嵌入，同时向输入中添加 cls token，以观察这将如何影响输出。
- **Ours 9**：我们没有让后续帧关注较早的帧，而是应用注意力掩码，确保每个帧只能关注位于其之前的帧，以保留序列的时间顺序。

1 结果

定量结果
表1 展示了 $JAAD_{all}$ 和 $JAAD_{beh}$ 数据集上的定量结果。这些表格将我们提出的模型与基准模型进行了比较：ConvLSTM[5], CAPformer[9], PCPA[3] & C3D[4]。在两个数据集中，所提出的模型在准确率、曲线下面积和 F1分数方面均取得了最佳成绩，而这些指标是二分类中最重要的指标。

表2 展示了在 $JAAD_{all}$ 和 $JAAD_{beh}$ 数据集上消融研究的定量结果。

定性结果
图2 展示了定性结果，其中我们使用一些现实场景的预录视频，将我们的模型与 PCPA 进行了比较。在图2所示的示例中，与 PCPA 不同，我们的模型能够正确预测真实意图。仔细观察这些示例后，我们推测这是因为我们的模型使用局部周边和全局上下文输入进行训练，因此对于目标行人周围的环境更加敏感；此外，由于局部周边输入中的目标行人被灰显，模型在目标行人是否朝向自车方面也更为稳健。实验的详细结果可通过以下媒体链接查看：https://youtu.be/CssklQmzP6c?si=93LcZ9fbboFE4i3Q。

图3 展示了我们在混乱环境中集成模型（Ours 7）的输出示例，其中行人已经走上道路，汽车则随处可见。各帧之间间隔 1–2 秒。该场景与模型训练所用的 JAAD 数据集中的几乎所有场景都不同，这表明我们的模型可应用于现实生活场景。实验的详细结果请见媒体文件https://youtu.be/FKbzKw1TSDc?si=60jQNtF-G2zaxNce。

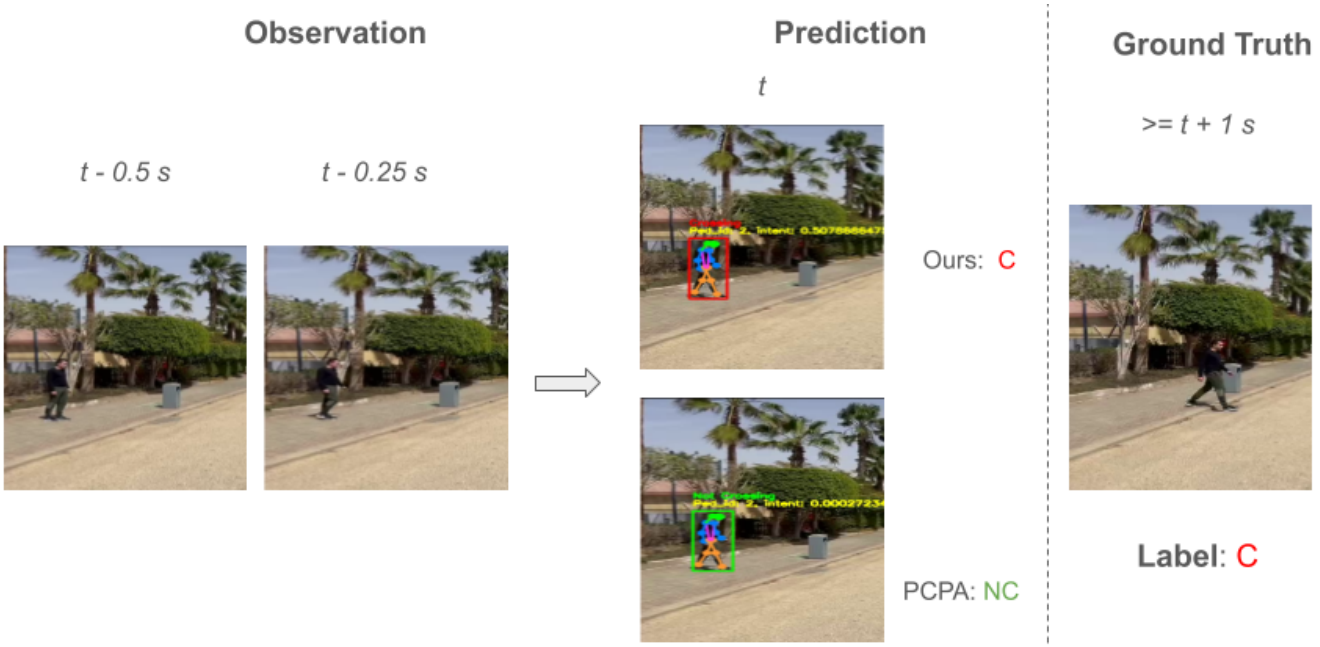

图2. 与 PCPA 不同，我们的模型能够正确预测行人的意图。红色边界框表示行人即将过马路，绿色表示不会过马路。



图3. 混乱环境。各帧之间间隔 1–2 秒。红色边界框表示行人即将过马路或正在过马路，绿色表示不会过马路。

参考文献

1. Kotseruba, I., Rasouli, A. & Tsotsos, J. K. 自动驾驶中的联合注意力（JAAD）。arXiv 预印本 arXiv:1609.04741 (2016)。2. Kotseruba, I., Rasouli, A. & Tsotsos, J. K. 他们想过马路吗？理解行人意图以进行行为预测。载于 2020 年 IEEE智能车辆研讨会（IV），1688–1693（IEEE，2020）。3. Kotseruba, I., Rasouli, A. & Tsotsos, J. K. 用于评估行人动作预测的基准。载于 IEEE/CVF冬季计算机视觉应用会议论文集，1258–1268（2021）。4. Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. 利用3D卷积网络学习时空特征。arXiv 电子预印本。arXiv 预印本 arXiv 1412 (2014)。5. Shi, X. 等。卷积长短期记忆网络：一种用于降水临近预报的机器学习方法。《神经信息处理系统进展》28 (2015)。6. Vaswani, A. 等。注意力就是一切。《神经信息处理系统进展》30（2017）。7. Dosovitskiy, A. 等。一张图像胜过 16x16 个词：面向大规模图像识别的 Transformer。arXiv 预印本 arXiv:2010.11929 (2020)。8. Arnab, A. 等。ViViT：视频视觉 Transformer。载于 IEEE/CVF国际计算机视觉会议论文集，6836–6846（2021）。9. Lorenzo, J. 等。CAPformer：使用 Transformer 进行行人过街动作预测。《Sensors》21，5694 (2021)。

10. Rasouli, A. 与 Kotseruba, I. Pedformer：通过跨模态注意力调制和门控多任务学习进行行人行为预测。载于 *2023*年*IEEE*国际机器人与自动化会议 *(ICRA)*，9844–9851（IEEE，2023）。**11.** Lorenzo, J.、Parra, I. 与 Sotelo, M. Intformer：借助 Transformer架构预测行人意图。*arXiv* 预印本 *arXiv:2105.08647*（2021）。**12.** Azarmi, M.、Rezaei, M.、Wang, H. 与 Glaser, S. Pip-net：真实环境中的行人意图预测。*arXiv* 预印本 *arXiv:2402.12810*（2024）。**13.** Gorishniy, Y.、Rubachev, I.、Khrulkov, V. 与 Babenko, A. 重新审视用于表格数据的深度学习模型。*Adv. Neural Inf.Process.Syst.* **34**，18932–18943（2021）。

作者贡献声明

A.R.E. 和 K.M.G. 设计了实验，A.R.E. 和 K.M.G. 开展了实验，所有作者均分析了结果。所有作者均审阅了稿件。

利益冲突

作者声明不存在任何利益冲突。

经费

本研究未获得任何组织的资助。

数据可用性

本研究期间生成和/或分析的数据集可在 JAAD网站获取：https://data.nvision2.eecs.yorku.ca/JAAD_数据集/

补充信息