针对人群流量预测模型的一致、有效且物理可实现的对抗攻击

Hassan Ali[1], Muhammad Atif Butt[1],Fethi Filali[2],Ala Al-Fuqaha[3],和 Junaid Qadir[41] 巴基斯坦旁遮普省拉合尔信息技术大学（ITU）[2] 卡塔尔多哈卡塔尔移动创新中心（QMIC）[3] 卡塔尔多哈哈马德·本·哈利法大学（HBKU）[4] 卡塔尔多哈卡塔尔大学

arXiv:2303.02669v1 [cs.LG] 5 Mar 2023

摘要——**近期研究表明，深度学习（DL）模型能够有效学习全市人流模式，从而助力更有效的城市规划和智慧城市管理。然而，众所周知，DL模型容易受到不易察觉的对抗扰动影响，导致性能不佳。尽管许多研究普遍探讨了此类对抗扰动，但深度学习人流预测模型的对抗脆弱性尤其尚未得到充分研究。本文对多种威胁设定下基于DL的人流预测模型的对抗脆弱性进行了严格分析，并作出三方面贡献。（1）我们通过正式定义人流预测输入的两种新属性——一致性和有效性，提出了*CaV-detect*，使其能够以0%的误接受率（FAR）检测标准对抗性输入。（2）我们利用通用对抗扰动和自适应对抗损失，提出自适应对抗攻击，以绕过*CaV-detect*防御。（3）我们提出*CVPR*，即一种一致、有效且物理可实现的对抗攻击，在扰动生成机制中明确引入一致性与有效性先验。我们发现，尽管人流预测模型容易受到对抗扰动的影响，但在物理场景中模拟这些扰动极具挑战性，尤其是在启用*CaV-detect*的情况下。我们还表明，*CVPR*攻击在误接受率和对抗损失指标方面显著优于经过自适应修改的标准攻击。最后，我们总结了本研究带来的有益见解，并指出了值得探索的未来研究方向。**

索引词—**深度神经网络、人流预测、对抗机器学习**

一、引言

人流预测问题旨在根据先前多个时间点的人流状态，预测未来某个时间点的人流状态。人流预测模型可广泛应用于众多领域，包括建模以及理解人类行为 [1]；交通管理 [2]；和智慧城市规划 [3]。尤其是在智慧城市场景中，人群流量预测模型对于识别过度拥挤区域至关重要，还可用于采取适当的预防性措施，确保人身安全。

深度神经网络（DNN）是解决人群预测问题的一种颇具前景的技术，已在许多复杂的实际任务中展现出其多功能性 [4]–[8]。近期研究成功地利用深度神经网络，不仅预测未来的人流状态 [2], [3], [9]，还用于相关任务，例如预测行人轨迹 [1] 和车辆行程时间。然而，深度神经网络的性能高度依赖其训练数据，因此容易受到对抗扰动的影响——这种难以察觉的噪声被刻意注入输入中，以改变深度神经网络输出 [7], [10], [11]。尽管已有研究提出了基于多种架构的人群流量预测模型，但据我们所知，这些模型的对抗脆弱性仍未得到充分探索。

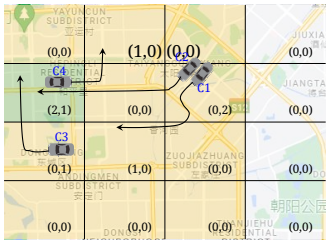


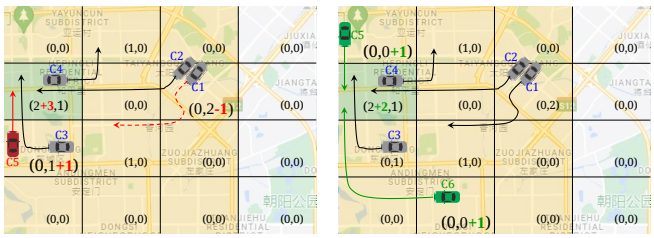
图1：区域划分网格，以及根据相邻区域（网格点）之间的人群流动计算流入矩阵和流出矩阵的示意图。我们通常假设邻域内存在邻接关系——绿色高亮区域的相邻网格点以黄色高亮显示。

本文通过研究三种流行且类型多样的人流预测模型在多种攻击设置（包括真实物理场景）下的最坏情况性能，弥补了这一空白：多层感知机（MLP） [3], 时空残差网络（STResnet） [2] 和时序图卷积神经网络（TGCN） [12]。评估时，我们采用TaxiBJ数据集，这是人流预测中最常用的数据集之一。TaxiBJ将一座城市划分为 32×32 个网格点（区域），并按半小时 间隔 [2], [3], [13], 记录区域级人群流入量[1] 和流出量[2] ，如图1所示。

挑战： 首先，不同人流预测模型的输入结构差异显著，因此，对抗评估结果可能无法公平地比较不同模型。例如，TGCN [12]仅接受以半小时间隔采集的预定义长度的人流状态历史作为输入。相比之下，时空残差网络 [2] 采用三

通讯作者：Junaid Qadir (junaid.qadir@qu.edu.qa)

[1]从相邻网格点流入某个网格点的设备总数。[2]从某个网格点流向相邻网格点的设备总数。



(a) 经 PGD攻击扰动　　(b) 经计算机视觉与模式识别会议 攻击扰动

图2：标准PGD攻击和我们提出的无效与有效对抗性输入的示意图，这些输入分别由两种方法生成。在(a)中，对于以绿色标出的网格点-(1，0)，记录到的总（扰动后）流入量为5。流入的5台设备中，有3台可能是从相邻区域（以黄色标出）流出的C2、C3和C5。另有两台设备是从哪里流出的？在(b)中，*CVPR*攻击根据网格点-(1，0)的流入量扰动，对相邻区域的流出量进行扰动，以保证物理合理性。

组输入，分别表示预定义长度的小时历史、日历史和周历史。为了公平评估不同架构所支持特征的鲁棒性，应在相似的输入设置下对模型进行评估。

其次，近年来，试图欺骗深度神经网络的攻击者与试图防御这些攻击的防御者之间展开了军备竞赛——在提出后的几个月内，大多数攻击和防御分别被更具自适应性的防御和攻击证明无效。因此，制定新颖的自适应防御和攻击策略，以全面分析针对深度神经网络的对抗攻击（及其局限性），既极其重要，也极具挑战性 [14], [15]。

研究发现与贡献： 我们首先分析了不同人流预测模型在三种标准对抗攻击——快速梯度符号法、迭代快速梯度符号法和PGD攻击——下的表现，并发现与其他深度学习模型类似，在多种设计选择下，人流预测模型也极易受到对抗攻击的影响。不过，我们注意到，这些脆弱性主要局限于数字攻击设置；该设置假定攻击者拥有对人流预测模型数字输入流水线的访问权限，并采取最坏情况攻击策略。

随后，我们指出自然人流输入必须满足两个属性——一致性 和有效性。尽管这些属性自然且直观，但据我们所知，先前的研究并未强调或使用它们。一致性 属性要求某一时间 t的人群流状态历史必须与之前各个时间的人流状态相一致。就有效性而言，根据定义，任意给定时间某个网格点的流入量和流出量，必须始终分别小于相邻网格点流出量和流入量的累计值。如图2(a)的示例所示，标准攻击产生的对抗扰动与这些关系相矛盾，因此在测试时很容易被判定为无效。我们注意到，标准攻击生成的对抗性输入不一致 且无效，并通过以下方式展示了这些属性的实用性：我们提出***CaV-detect***，这是一种新颖的一致性与有效性检查机制，可在测试时通过分析流入量和流出量矩阵（检查有效性）并比较人群流状态历史（确保一致性）来检测对抗性输入。结果表明，*CaV-detect* 能够以0%的误报率检测标准对抗性输入（漏报率 $\leq$ 0.5%）。

假设攻击者是专业攻击者，我们通过结合通用对抗扰动 [16] 和自适应对抗损失，自适应地修改标准对抗攻击，以规避*CaV-detect* 。与误报率为0%的非自适应标准攻击相比，自适应攻击通常达到 $\geq$80%的误报率（漏报率为$\leq$ 0.5%）。

随后，我们提出计算机视觉与模式识别会议 **攻击，这是一种一致、有效且物理可实现的**对抗攻击，它在对抗输入生成机制中明确引入一致性与有效性先验，以寻找一致且有效的对抗扰动（见图2(b)），并且在误报率（≈100%）和针对*CaV-detect*的对抗损失两方面均优于标准攻击和自适应攻击。

我们还注意到，标准对抗攻击生成的扰动通常为负值，因此无法由物理攻击者实现——物理攻击者只能控制城市中数量有限的物理设备，且无法访问数字输入流水线。因此，我们修改了对抗损失，通过生成能够在现实世界中进行物理模拟的对抗扰动，提出并评估了一种物理上可实现的攻击。我们的研究结果表明，在物理场景下实现对抗扰动需要数量大得不切实际的对抗性控制设备，尤其是在启用 *CaV-detect* 时。

最后，我们的定性评估表明，人流预测模型具有有限表达能力——尽管所得模型表现出较小的测试误差，但无法生成某些输出。我们将其归因于 TaxiBJ 数据包含聚类且高度相似的人流状态 [3]。我们的主要贡献如下：

- 我们首次研究了人流预测模型的对抗脆弱性。
- 我们形式化地提出了人流预测输入的两项新性质——一致性和有效性，并提出一种名为*CaV-detect* 的新型防御方法，以证明这两项性质的实用性；该方法的误报率为 0%，错误拒绝率（FRR）为 $\leq$0.5%。
- 我们将自适应损失与通用对抗扰动相结合，以全面测试*CaV-detect*。
- 我们将一致性与有效性先验引入对抗性输入生成机制，提出计算机视觉与模式识别会议攻击，以解决标准攻击的若干缺陷。

II. 相关工作

由于智能交通系统（ITS）的近期发展，道路交通拥堵预测正逐渐成为减少交通管理中延误及相关成本的关键步骤之一 [17]。下文中，

我们重点介绍一些近期具有代表性的群体流状态预测研究。

A. 人群流状态预测

根据数据的特性、结构和质量，人们会采用各种机器学习（ML）技术来构建道路交通拥堵模型。在文献中，这些人流预测技术通常分为三个主要类别——基于概率与统计推理的人流模型 [18]–[30], 浅层机器学习技术 [31]–[39] 以及深度学习（DL）模型 [17],[24], [40]–[43]。我们的工作着重研究基于深度学习的人流状态预测模型的对抗脆弱性。具体而言，我们选择了三个架构显著不同的人流预测模型进行鲁棒性评估。选择依据包括架构的新近性、多样性、与问题的相关性以及受欢迎程度。所有这些模型架构均在其原始论文中使用TaxiBJ数据集进行了训练和评估。

MLP架构。 在近期的工作中，Jiang等人 [3]提出了2021年的TaxiBJ数据集，并使用一个简单的多层感知机模型作为基准来评估其结果。我们选择多层感知机模型，是因为它较新颖、简单且具有对抗迁移性——近期研究表明，与其他架构相比，针对多层感知机模型生成的对抗性输入具有更强的通用性，并且能更有效地迁移到不同架构上 [5]。

STResnet架构。 Zhang等人提出的STResnet模型 [2]基于时空残差单元构建，通过卷积层建模空间依赖关系，并将近期的人群流动状态拼接成一个元组，以建模时间依赖关系。我们选择该模型，是因为其架构（具有时空特性）与本研究高度相关，且颇受欢迎（引用次数为 1425 次[3]）。然而， [2] 中提出的模型是在将小时、日和周历史拼接成单个输入元组后进行训练的。为公平比较不同架构，我们调整其架构，使之与所有架构通用的输入结构相匹配，仅使用半小时历史进行训练。

T-GCN架构。 在人群流动模式建模中，图卷积网络（GCN）是公认的首选方法。Zhao等人 [12]提出了一种时序图卷积网络（时空图卷积网络），利用随时间更新的卷积图网络，对输入中的时间依赖关系和空间依赖关系进行建模。我们选择其提出的时空图卷积网络模型，是因为该模型与本研究高度相关（具有时空架构），且颇受欢迎（引用次数为 839 次[4]）。

B. 针对深度学习模型的对抗攻击

对抗攻击是对输入进行微小且不可察觉的改变，以欺骗深度神经网络。自从 Szegedy等人发现图像分类中的这一现象以来 [44]，大量研究表明，在包括计算机视觉（CV）在内的各种应用中，深度神经网络通常容易受到此类攻击，这些应用还包括音频处理 [45], 、网络 [46]–[48]和自然语言处理（NLP） [5], [49]。对抗攻击可能采用黑盒 [7], [50], [51] 或白盒威胁模型 [52]–[54]。尽管人们已采取多种措施来防御这些攻击 [52], [55]–[58], ，认证防御 [59] 和对抗训练 [53] 是仅有的能够为深度学习模型提供可靠鲁棒性的有效防御方法 [14], [60]。

在这项工作中，我们假设采用白盒威胁模型，即攻击者了解模型架构和权重。设输入为 $x \in \mathcal{X}$，其中 $\mathcal{X}$ 表示有效输入特征空间，并产生真实输出 $\mathcal{F}_\theta(x)$，其中 θ 表示 $\mathcal{F}$ 的可学习参数。攻击的目标是计算对抗扰动 δ^*，以便在将扰动添加到输入后，从模型获得期望输出 y_{target}。

$$\delta^* = \underset{\delta \in \mathcal{B}(\epsilon)}{\operatorname{argmin}} (\mathcal{F}_\theta(x+\delta) - y_{target})^2 \tag{1}$$

其中， $\mathcal{B}(\epsilon)$ 表示预先定义的允许扰动有界集合。对于 $\mathcal{B}(\epsilon)$, 最常见的选择之一是 l_∞ 球，定义为： $\boldsymbol{\delta} \in \mathcal{B}_\infty(\epsilon)$: $= -\epsilon \leq \boldsymbol{\delta} \leq \epsilon$。方程（26）根据攻击算法 [11] 进行迭代优化。

快速梯度符号法(FGSM) [52]: FGSM攻击通过单步计算对抗扰动 δ^* 来实现式(1)，即:

$$\delta^* = \delta - \epsilon \times \operatorname{sign}\left(\frac{\partial (\mathcal{F}_\theta(x+\delta) - y_{target})^2}{\partial \delta}\right) \tag{2}$$

迭代式FGSM: i-FGSM攻击通过在 N 步中计算对抗扰动 δ^* 来实现式(1)，即:

$$\delta_{i+1} = \delta_i - \frac{\epsilon}{N} \times \operatorname{sign}\left(\frac{\partial (\mathcal{F}_\theta(x+\delta) - y_{target})^2}{\partial \delta_i}\right)$$
$$\delta_{i+1} = \operatorname{clip}(\delta_{i+1}, \epsilon, \epsilon) \tag{3}$$

其中， N 为迭代次数， $\delta_0 = \delta$，且 $\delta^* = \delta_N$。

投影梯度下降(PGD) [53]: 通过重复执行式(4)中的步骤 $[0, N–1]$ 步，可实现式(1)中的优化，

$$\delta_{i+1} = \delta_i - \alpha \times \operatorname{sign}\left(\frac{\partial (\mathcal{F}_\theta(x+\delta) - y_{target})^2}{\partial \delta_i}\right)$$
$$\delta_{i+1} = \operatorname{clip}(\delta_{i+1}, \epsilon, \epsilon) \tag{4}$$

其中， $\delta_0 = \delta$、 $\delta^* = \delta_N$和 α 表示PGD步长。注意， α 是PGD攻击的可调超参数。

III. 方法论

我们首先以 TaxiBJ 数据集为背景，对人流预测问题进行建模，并正式定义人群流状态输入的一致性和有效性属性。基于这些属性，我们提出 *CaV-detect*，通过分析输入的一致性和有效性来检测经过对抗扰动的输入。最后，我们提出用于生成一致、有效且物理上可实现的对抗扰动的新算法 *CVPR*攻击。

[3]scholar.google.com，3 2022年 4 November
scholar.google.com，3 November 2022年

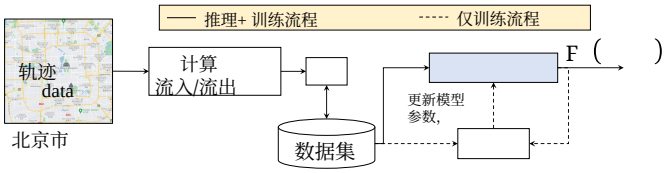


图3：展示了用于 TaxiBJ数据集的人流预测模型训练设置。首先，将从城市收集的轨迹数据转换为流入/流出矩阵，并使用 $\mathcal{T}$ 进行变换，然后将其保存到内存中，并与历史集合拼接，形成输入人流预测模型的元组输入 $\mathbf{X}_h(t) = \bigcup_{i=0}^{h} \mathbf{x}^{t-i}$。

A. 人群流状态预测

问题建模。 本研究使用的公开 TaxiBJ 人群流数据集 [3] 将城市划分为大小为 $l_1 \times l_2$ 的二维网格，其中每个网格点在现实中覆盖 1000 平方米的区域。该数据集每隔 30 分钟以流入量和流出量矩阵组成的元组形式报告全市人流量。在任意给定时间 t，整数人流状态 $\mathbf{n}^t \in \mathbb{Z}^{2\times l_1 \times l_2}$定义为由流入量和流出量矩阵构成的元组，分别记为 $\mathbf{n}_{in}^t \in \mathbb{Z}^{l_1 \times l_2}$ 和 $\mathbf{n}_{out}^t \in \mathbb{Z}^{l_1 \times l_2}$，用于表示在城市网格中流入和流出各网格点的设备数量（$\approx$ 人员 [2]）。形式化地，

$$\mathbf{n}^t = (\mathbf{n}_{in}^t, \mathbf{n}_{out}^t) \tag{5}$$

我们将人流状态定义为$\mathbf{x}^t \in \mathbb{R}^{2\times l_1 \times l_2}$,

$$\mathbf{x}^t = (\mathbf{x}_{in}^t, \mathbf{x}_{out}^t) = \mathcal{T}(\mathbf{n}^t) = (\mathcal{T}(\mathbf{n}_{in}^t), \mathcal{T}(\mathbf{n}_{out}^t)) \tag{6}$$

其中， $\mathcal{T}(\cdot)$表示逐元素（某种程度上）可逆的变换函数。通常的做法是选择 $\mathcal{T}$，使得 $\forall \mathbf{n}^t \in [0..\infty], \mathbf{x}^t \in [0,1]$。遵循先前研究的做法 [2], [3], , 我们在实验中使用 $\mathcal{T}(\mathbf{n}^t) = \min(\mathbf{n}^t/1000, 1)$。

我们的目标是学习一个模型 $\mathcal{F}_\theta$，在给定当前状态和先前状态 $\bigcup_{i=0}^{h} \mathbf{x}^{t-i}$的情况下，预测紧接着的未来人流状态 $t+1$。未来将其记为$\mathbf{X}_h(t)$，其中 h 是历史长度，表示将总数为 h 的先前人流状态与当前状态拼接在一起，作为元组输入提供给 $\mathcal{F}_\theta$。形式化地，

$$\mathbf{y}^{t+1} = \mathcal{F}_\theta(\mathbf{X}_h(t)) \tag{7}$$

其中$\mathbf{y}^{t+1}$ 表示模型的输出。与先前研究类似，我们将上述问题作为回归任务来求解，学习一个模型 $\mathcal{F}_{\theta^*}$，使其在整个TaxiBJ数据集 $\mathcal{D}$上最小化输出与真实值之间平方差的期望，

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{x\sim\mathcal{D}}[(\mathcal{F}_\theta(\mathbf{X}_h(t)) - \mathbf{x}^{t+1})^2] \tag{8}$$

我们用于训练人群流量预测模型的训练设置 $\mathcal{F}_\theta$如图3所示。

人流状态输入的属性。我们在此正式定义人流预测输入的两个关键属性：一致性和有效性，这些属性促成了*CaV-detect*的开发。我们还对公式（1）进行了正式分析（特别针对上述属性），以突出对人流预测输入实施对抗攻击的局限性。

1) 一致性： 我们考虑一段人群流动状态序列， $\bigcup_{i=0}^{-2h} \mathbf{x}^{t-i}$，这些状态记录于从 $t-2h$ 到 t的不同时间间隔，

$$\bigcup_{i=0}^{2h} \mathbf{x}^{t-i} = \{\mathbf{x}^{t-2h}, \ldots, \mathbf{x}^{t-h}, \ldots, \mathbf{x}^{t}\} \tag{9}$$

其中， h 表示历史长度，$\mathbf{x}^t$ 表示时间 t的人群流动状态。数据预处理步骤会构建一个包含 h 个先前人群流动状态的历史集合， $\bigcup_{i=1}^{h} \mathbf{x}^{t-i}$，并将该历史集合与当前人群流动状态 $\mathbf{x}^t$拼接，形成输入人流预测模型的元组输入 $\bigcup_{i=0}^{h} \mathbf{x}^{t-i}$。

我们注意到，对于 $1 \le k \le h$，时间 t的历史集合是历史集合 $\bigcup_{i=1}^{l_h-k} \mathbf{x}^{(t-k)-i}$的一个子集与模型在时间 $t-k$的输入中的人群流动状态 $\mathbf{x}^{(t-k)}$的并集。当且仅当以下条件成立时，输入才是一致的:

$$\forall k \in [1..h], \bigcup_{i=0}^{h-k} \mathbf{x}^{(t-k)-i} = \bigcup_{i=k}^{h} \mathbf{x}^{t-i} \tag{10}$$

这引出了我们稍后提出的一致性检查机制。简单来说，如图4中的示例所示，任意时间 t 的历史集合都应与先前时间的人流状态保持一致。**备注**。 对于每个 t，标准对抗攻击都必须学习一组新的扰动 $\bigcup_{i=0}^{h} \boldsymbol{\delta}^{t-i}$，且该组扰动独立于（因此不同于）针对某个先前时间 $t-k$ 学得的扰动 $\bigcup_{i=0}^{h} \boldsymbol{\delta}^{(t-k)-i}$。形式上，对于标准对抗攻击，

$$\forall k \in [1..h], \bigcup_{i=k}^{h} \boldsymbol{\delta}^{t-i} \neq \bigcup_{i=0}^{h-k} \boldsymbol{\delta}^{t-k-j} \tag{11}$$

简单来说，标准对抗攻击生成的对抗扰动（以及由此产生的对抗扰动输入）是不一致的，因此可以被我们之前形式化定义的*CaV-detect* 机制轻易检测到。

2) 有效性：考虑图1所示的一个 4×4 网格。对于一个网格点-(0，1)（绿色阴影区域），从其相邻网格点（黄色阴影区域）进入该网格点的设备总数为2（C2和C3）。由于这些设备必须从相邻网格点流出，因此相邻网格点的总流出量至少必须为2。更一般地，我们将邻接关系定义为图1所示网格点的 2^{nd} 邻域范围内。给定一个特定的网格点-(p_1， p_2)，令A_n(p_1， p_2)表示 n^{th} 邻域中与网格点-(p_1， p_2)相邻的网格点集合，

$$A_n(p_1, p_2) = \bigcup_{i=-n}^{n} \bigcup_{\substack{j=-n \\ |i|+|j|\neq 0}}^{n} (p_1 - i, p_2 - j) \tag{12}$$

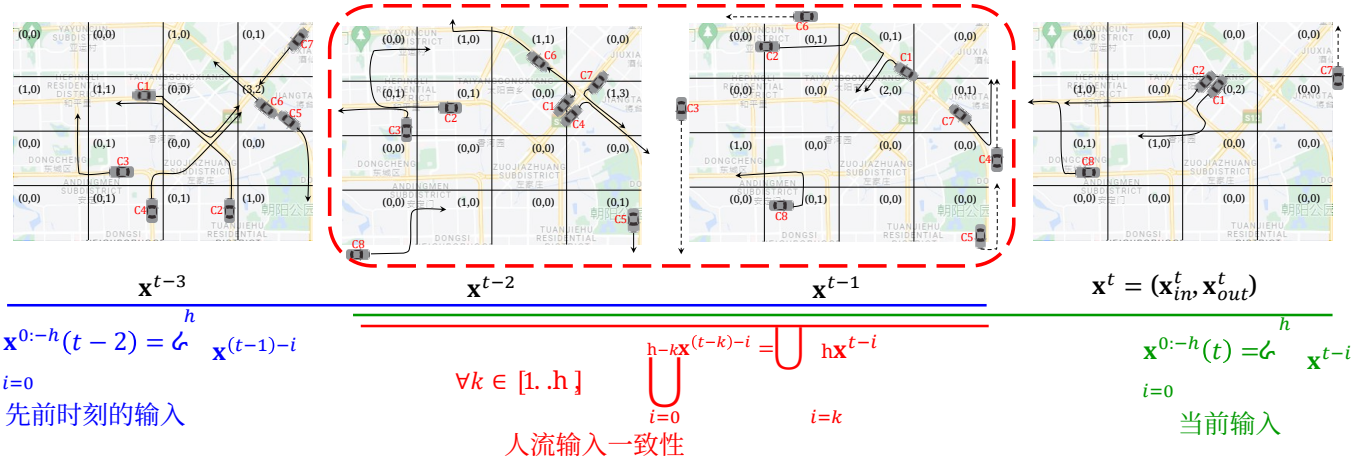


图4：人流状态输入一致性属性的示意图。在任意时间 t，人群流状态历史都必须与之前时刻记录的人流状态一致。

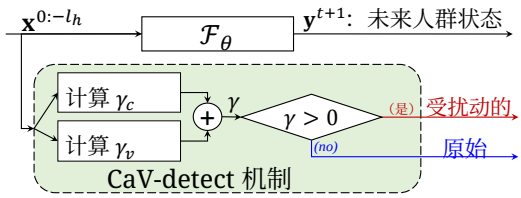


图5：展示我们新提出的 CaV-detect 方法，该方法与人流预测模型集成，以便在运行时检测对抗性输入。对于任意输入，CaV-detect 会检查输入的一致性 $\gamma_c > 0$ 和有效性 $\gamma_v > 0$。如果任一项检查未通过，就会将该输入标记为对抗性输入。*CaV-detect* 无需重新训练模型，并且可以与现成模型集成。

其中 n 是用于确定邻接关系的邻居数量。根据定义，在任意给定时间 t，网格点-(p_1，p_2)的流入量是从其相邻网格点$A_n(p_1$，p_2)进入该网格点的设备总数。因此，从 $A_n(p_1$，p_2)流出的总量必须至少等于进入(p_1，p_2)的总流入量。令$\mathbf{x}^t_{in}(p_1$，p_2)和$\mathbf{x}^t_{out}(p_1$，p_2)分别表示时间 t时网格点-(p_1，p_2)的流入量和流出量。只有当输入人流预测模型的输入满足有效条件时，

$$\mathbf{x}^t_{in}(p_1,p_2) \le \sum_{(p'_1,p'_2)\in A_n(p_1,p_2)} \mathbf{x}^t_{out}(p'_1,p'_2) \quad (13a)$$

$$\mathbf{x}^t_{out}(p_1,p_2) \le \sum_{(p'_1,p'_2)\in A_n(p_1,p_2)} \mathbf{x}^t_{in}(p'_1,p'_2) \quad (13b)$$

说明。 在生成对抗扰动时，式（1）中形式化定义的标准对抗攻击并不遵循流入矩阵与流出矩阵之间的这种关系，因此可以在运行时被检测到。

B. CaV-detect：检测对抗扰动输入的一致性与有效性检查机制

在此，我们利用先前定义的人流输入的两个属性，提出 *CaV-detect*，这是一种新颖的输入验证机制，用于检测针对人流预测模型的对抗扰动输入。总而言之，我们的 *CaV-detect* 方法论包含两个主要步骤——一致性检查机制和 有效性 检查机制。如果模型输入未能通过上述任一项检查，则认为该输入经过了对抗性扰动。*CaV-detect* 的具体步骤如下。

一致性检查机制： 令 $\mathbf{x}^t$ 表示任意给定时间 t 的人流状态。于是，人流预测模型的输入 $\mathcal{F}_{\theta^*}$ 可表示为 $\mathbf{X}_h = \bigcup_{i=0}^{h} \mathbf{x}^{t-i}_h$ (t)

$\mathbf{x}$，其中 表示历史长度。我们的一致性检查机制分两步进行：1) 首先，我们将之前各个时间 $t-k$ 收到的所有模型输入 $\mathbf{x}^{(t-k)}$ 保存到内存 $\forall k \in [1..h]$ 中。2) 注意到，之前各个时间 $t-k$ 收到的模型输入会再次出现在当前时间 t 收到的输入的历史集合中，因此我们计算不同时刻经过适当裁剪的模型输入之间的差异

$$\gamma_c = \sum_{k=1}^{h} \left| \bigcup_{i=k}^{h} \mathbf{x}^{t-i} - \bigcup_{i=0}^{h-k} \mathbf{x}^{(t-k)-i} \right| \ge 0 \quad (14)$$

其中，γ_c 是不一致性得分—— γ_c 越接近零，输入的一致性越高。

3) 如果 $\gamma_c > 0$，则将该输入标记为对抗性输入。

有效性检查机制： 令 $\mathbf{x}^t = (\mathbf{x}^t_{in}, \mathbf{x}^t_{out})$ 表示任意给定时间 t 的人流状态，其中 $\mathbf{x}^t_{in}, \mathbf{x}^t_{out} \in \mathbb{R}^{l_1 \times l_2}$。我们的有效性检查机制分为以下四个步骤。

1) 我们首先定义一个滤波器 $\mathbf{f} \in \mathbb{Z}^{5\times5}$，使得 $\forall p_1, p_2 \in [0..4]$，

$$\mathbf{f}(p_1,p_2) = \begin{cases} 1, & p_1 = 2, p_2 = 2 \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

2) 其次，我们通过同时分析输入中的流入量和流出量矩阵，分别计算流入量和流出量无效性分数，记为 γ_{v_i} 和 γ_{v_o}。

$$\gamma_{v_i} = \mathbf{x}_{in}^{t-i} \circledast \mathbf{f} - \mathbf{x}_{out}^{t-i} \circledast (\mathbf{1} - \mathbf{f}) \leq \mathbf{0} \quad (16)$$

$$\gamma_{v_o} = \mathbf{x}_{out}^{t-i} \circledast \mathbf{f} - \mathbf{x}_{in}^{t-i} \circledast (\mathbf{1} - \mathbf{f}) \leq \mathbf{0} \quad (17)$$

其中，$\circledast$表示二维卷积操作。

3）最后，我们基于步骤2中计算得到的流入量和流出无效性分数，计算输入无效性分数 γ_v。

$$\gamma_v = \text{relu}(\gamma_{v_i} + \gamma_{v_o}) \quad (18)$$

其中，relu表示深度学习文献中常用的修正线性单元函数。

4）若 $\gamma_v > 0$，则该输入被标记为对抗性输入。

请注意，*CaV-detect*采用的两种检查机制均为模型无关。因此，*CaV-detect* 可与具有不同架构的预训练人群流预测模型结合使用，而不会削弱其效能。

C. CVPR-attack：针对人流预测模型的一致、有效且物理可实现的对抗攻击

鉴于前文已形式化阐述标准对抗攻击在实际应用中的局限性，本节提出计算机视觉与模式识别会议攻击——一种具有一致性、有效性 且物理上可实现的对抗攻击。在任意给定时间 t，我们考虑模型 $\mathcal{F}_\theta$ 的输入 $\mathbf{X}_h(t) = \bigcup_{i=0}^{h} \mathbf{x}^{t-i}$。我们的目标是对输入添加扰动 $\boldsymbol{\Delta}_h(t) = \bigcup_{i=0}^{h} \boldsymbol{\delta}^{t-i}$，使模型输出更接近对抗目标 $\mathbf{y}_{target}^{t+1}$。

1) **一致性：** 为确保扰动的一致性，我们利用通用对抗扰动来约束 $\boldsymbol{\Delta}_h(t)$，使得 $\forall i \in [0..h], \boldsymbol{\delta}^{t-i} = \boldsymbol{\delta}^u$。

$$\boldsymbol{\Delta}_h(t) = \bigcup_{i=0}^{h} \boldsymbol{\delta}^u = \bigcup_{i=0}^{h} (\boldsymbol{\delta}_{in}^u, \boldsymbol{\delta}_{out}^u) \quad (19)$$

2) **有效性：** 为确保有效性，我们提出一种新机制：给定扰动流入矩阵 $\boldsymbol{\delta}_{in}^u$，生成扰动流出矩阵 $\boldsymbol{\delta}_{out}^u$。具体而言，给定 $\boldsymbol{\delta}_{in}^u$、一个特定网格点 (p_1, p_2)，以及其 n^{th} 邻域中的一组相邻网格点 $A_n(p_1, p_2)$，我们学习一个扰动分布矩阵 $\mathbf{W} \in \mathbb{R}^{l_1 \times l_2 \times (2n+1)^2 - 1}$，首先将流入 (p_1, p_2) 的总扰动分配到 $A_n(p_1, p_2)$ 之间，

$$\boldsymbol{\delta}_{out}^* = \boldsymbol{\delta}_{in}^u \odot \frac{\sigma(\mathbf{W})}{\sum_i \sigma(\mathbf{W})[:, :, i]} \quad (20)$$

其中，$\odot$ 表示逐元素（Hadamard）乘法，$\boldsymbol{\delta}_{out}^* \in \mathbb{R}^{l_1 \times l_2 \times (2n+1)^2 - 1}$ 表示一组分布式扰动流出矩阵，用于 $\boldsymbol{\delta}_{out}^*$ 满足人流输入的有效性条件。随后，通过累加相关的分布式流出量，计算（p_1，p_2）的总扰动流出量，

$$\boldsymbol{\delta}_{out}^u(p_1, p_2) = \sum_{i=-n}^{n} \sum_{\substack{j=-n \\ |i|+|j| \neq 0}}^{n} \boldsymbol{\delta}_{out}^*(p_1 - i, p_2 - j, k) \quad (21)$$

其中，k 定义如下，

$$k = \begin{cases} (2n+1)(i+n) + j + n - 1, & i > 0, j > 0 \\ (2n+1)(i+n) + j + n, & \text{otherwise} \end{cases} \quad (22)$$

换言之，$\boldsymbol{\delta}_{out}$ 被计算为$\boldsymbol{\delta}_{in}$ 和$\mathbf{W}$ 的函数 f，如图 6 所示。随后，式（19）可重写为：

$$\boldsymbol{\Delta}_h(t) = \bigcup_{i=0}^{h} (\boldsymbol{\delta}_{in}^u, \boldsymbol{\delta}_{out}^u) = \bigcup_{i=0}^{h} (\boldsymbol{\delta}_{in}^u, f(\boldsymbol{\delta}_{in}^u, \mathbf{W})) \quad (23)$$

3) **物理可实现性：** 为了在任意给定时间 t改变模型输出，攻击者学习到的扰动 $\boldsymbol{\Delta}_h(t) = \bigcup_{i=0}^{h} \boldsymbol{\delta}^{t-i}$ 会因不同的 i而显著不同。实际上，这类攻击仅在数字攻击设置下可行。在物理攻击设置下实现此类攻击，需要攻击者精确控制每个网格点中的设备数量；这颇具挑战性，因为攻击者必须反复重新部署对抗设备，或者拥有足够数量的对抗设备并反复将其开启和关闭，以模拟 $\boldsymbol{\Delta}_h(t)$。通用对抗扰动通过为每个时间间隔生成单个最有效的扰动，自然地解决了这一问题。

此外，生成 $\boldsymbol{\delta}^u \in \mathcal{B}_\infty(\epsilon)$球仅适用于数字攻击设置。对于物理攻击设置，攻击者只能实现 $\boldsymbol{\delta} > 0$ 扰动（例如，通过实际增加一定数量的对抗设备）。因此，对于物理攻击，我们将扰动优化至 $\mathcal{B}_\infty(\mathbf{0}, \epsilon)$界内。

在生成扰动的过程中，$\boldsymbol{\delta}^u$，我们会迭代更新 $\boldsymbol{\delta}_{in}$ 和 $\mathbf{W}$，以寻找最优扰动。具体而言，我们优化以下损失函数，

$$\mathcal{L}_{upa}\left(\mathbf{X}_h(t), (\boldsymbol{\delta}_{in}^u, \mathbf{W})\right) = \left| \mathcal{F}\left(\mathbf{X}_h(t) + \bigcup_{i=0}^{h} \boldsymbol{\delta}^u\right) - \mathbf{y}_{target}^{t+1} \right|,$$

$$\boldsymbol{\delta}_{in}^u, \mathbf{W} = \underset{\boldsymbol{\delta}_{in}^u \in \mathcal{B}(\epsilon), \mathbf{W}}{\text{argmax}} \; -\mathcal{L}_{upa}\left(\mathbf{X}_h(t), (\boldsymbol{\delta}_{in}^u, \mathbf{W})\right) \quad (24)$$

其中，$\boldsymbol{\delta}^u = (\boldsymbol{\delta}_{in}^u, \boldsymbol{\delta}_{out}^u)$ 表示对所有 $\mathbf{x}^t \in \mathcal{D}$均保持不变的通用对抗扰动，$\epsilon$ 是前文所述的最大扰动预算。为了实现物理可实现性（限制 3），我们在使用梯度下降最大化 $\mathcal{L}_{adv}$ 的同时，反复裁剪负值，并将 $\boldsymbol{\delta}^u$ 投影到 $\mathcal{B}(\epsilon)$ 球内。详情见算法 1。

IV. 实验设置

A. 威胁模型

对于本文中的所有实验，我们假设采用白盒威胁模型，在该模型中，攻击者完全了解人流预测模型的架构及其学习到的参数，θ。此外，我们始终假设采用定向攻击场景，即攻击者的目标是对输入施加扰动，使模型输出实现特定的人群流状态，该状态称为目标状态，记为$\mathbf{y}_{target}^{t+1}$。其目标是

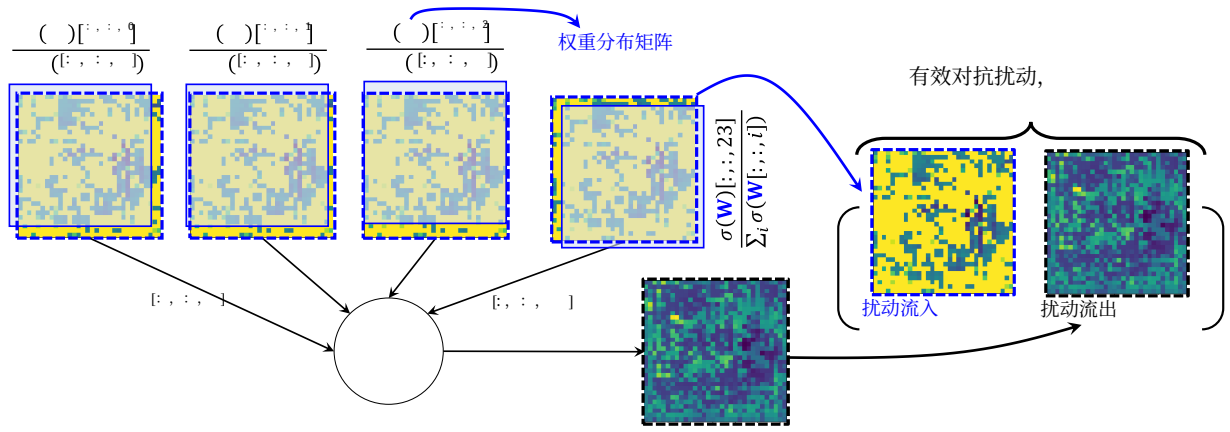


图 6：展示新提出的 计算机视觉与模式识别会议 攻击方法，该方法用于生成有效对抗扰动。以蓝色突出显示的术语表示攻击过程中为优化对抗损失而更新的变量。给定一个 扰动流入矩阵 $\boldsymbol{\delta}_{in}$，通过将 $\boldsymbol{\delta}_{in}$ 与经过适当移位的归一化扰动分布矩阵 $\mathbf{W}$ 逐元素应用，计算得到一组 分布式扰动流出矩阵 $\boldsymbol{\delta}^{*}_{out}$。最后，将 $\boldsymbol{\delta}^{*}_{out}$ 的所有切片相加，计算得到 扰动流出矩阵 $\boldsymbol{\delta}_{out}$。

算法 1 计算机视觉与模式识别会议攻击算法

输入：
$\mathbf{X}_h(t) \leftarrow$ 人群流状态历史$\mathcal{F} \leftarrow$ 训练模型$\mathbf{y} \leftarrow$ 输出人群流状态$N \leftarrow$# 次迭代$\epsilon \leftarrow$ 最大扰动预算

输出：
$\boldsymbol{\delta}^{u} \leftarrow (\boldsymbol{\delta}^{u}_{in},\ \boldsymbol{\delta}^{u}_{out}) \leftarrow$ 通用对抗扰动

1: 定义 $i \leftarrow 1,\ \boldsymbol{\delta}_{in} \leftarrow \mathbf{0},\ \mathbf{W} \leftarrow -\mathbf{5}$
2: 定义 $\eta \leftarrow (5 \times \epsilon)/N$
3: **重复**
4: $\quad \boldsymbol{\delta}^{u} \leftarrow (\boldsymbol{\delta}^{u}_{in},\ f(\boldsymbol{\delta}^{u}_{in},\ \mathbf{W}))$
5: $\quad \mathcal{L}_{adv} \leftarrow \left(\mathcal{F}\left(\mathbf{X}_h(t) + \bigcup_{i=0}^{h} \boldsymbol{\delta}^{u}\right) - \mathbf{y}^{t+1}_{target}\right)^{2}$
6: $\quad \boldsymbol{\delta}^{u}_{in} \leftarrow \boldsymbol{\delta}^{u}_{in} - \eta \times$ 符号$\left(\frac{\partial \mathcal{L}}{\partial \boldsymbol{\delta}^{u}_{in}}\right)$
7: $\quad \mathbf{W} \leftarrow \mathbf{W} - \eta \times$ 符号$\left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}}\right)$
8: $\quad \boldsymbol{\delta}^{u}_{in} \leftarrow$ 截断$(-\epsilon,\ \epsilon)$ —— (l_∞ 边界)
9: $\quad \boldsymbol{\delta}^{u}_{out} \leftarrow f(\boldsymbol{\delta}^{u}_{in},\ \mathbf{W})$
10: $\quad i \leftarrow i + 1$
11: **直到** $i \leq N$

白盒攻击者的目标是学习扰动 $\boldsymbol{\Delta}_h(t) \in \mathcal{B}(\epsilon)$，将其添加到输入后，可使其与攻击者定义的目标状态$\mathbf{y}^{t+1}_{target}$具有最大相关性。

$$\mathcal{L}_{adv}\left(\mathbf{X}_h(t), \boldsymbol{\Delta}_h(t)\right) = \left|\mathcal{F}(\mathbf{X}_h(t) + \boldsymbol{\Delta}_h(t)) - \mathbf{y}^{t+1}_{target}\right| \tag{25}$$

我们实验了三种不同的白盒威胁配置，具体如下所述，并在图7中进行了说明。

WB-blind 威胁模型： 在这种白盒威胁模型中，假设攻击者不知道流水线中部署的 *CaV-detect*机制。WB-blind 攻击者的目标形式化如下，

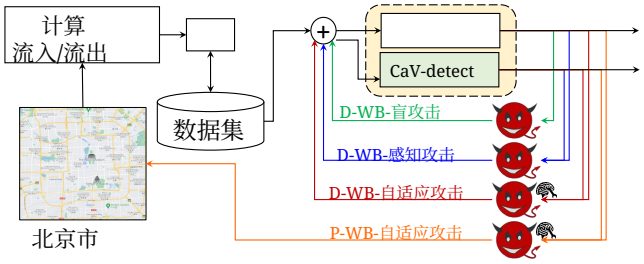


图7：我们实验中使用的不同白盒威胁模型的示意图与比较。D-WB 表示白盒威胁模型下的数字攻击设置。P-WB 表示白盒威胁模型下的物理攻击设置。自适应攻击者会调整扰动生成机制以欺骗*CaV-detect*。

$$\boldsymbol{\Delta}_h(t) = \underset{\boldsymbol{\Delta}_h(t) \in \mathcal{B}_\infty(\epsilon)}{\operatorname{argmax}} -\mathcal{L}_{adv}\left(\mathbf{X}_h(t), \boldsymbol{\Delta}_h(t)\right) \tag{26}$$

其中，$\mathcal{B}_\infty(\epsilon)$ 是一个 l_∞ 球，定义为，$\boldsymbol{\delta} \in \mathcal{B}_\infty(\epsilon) := -\epsilon \leq \boldsymbol{\delta} \leq \epsilon$。式 (26) 根据攻击算法 [11] 进行迭代优化。

白盒感知威胁模型： 该白盒威胁模型假设攻击者完全了解流水线中的*CaV-detect* 机制，并试图通过*CaV-detect* 逃避检测，同时努力在模型输出中产生目标状态。形式上，我们定义一个拉格朗日函数，

$$\boldsymbol{\Delta}_h(t) = \underset{\boldsymbol{\Delta}_h(t) \in \mathcal{B}(\epsilon)}{\operatorname{argmax}} -\mathcal{L}_{adv}\left(\mathbf{X}_h(t), \boldsymbol{\Delta}_h(t)\right) - \lambda \times (\gamma_c + \gamma_v) \tag{27}$$

其中 $\lambda = 10^{10}$ 是拉格朗日乘子。根据攻击算法对式（27）进行迭代优化。

注：我们在白盒感知威胁模型下进行实验，发现白盒感知攻击者无法计算出能够造成任何显著影响的对抗扰动

模型输出变化。我们认为，这是由于严格的一致性和有效性检查机制所致，该机制在优化式（27）时会导致相互矛盾的梯度更新。因此，我们未在论文中报告定量结果。

WB 自适应威胁模型： 与 WB 感知类似，该威胁模型同样假设攻击者完全了解流水线中的*CaV-detect* 机制，并试图通过*CaV-detect* 规避检测，同时显著影响输出，使其趋向目标状态，方法是自适应地修改攻击算法。更具体地说，自适应攻击者会修改攻击，使式（27）更容易优化。为此，我们的自适应攻击者利用通用对抗扰动算法，自然地规避一致性检查机制。于是，对抗损失函数可以定义为如下的拉格朗日函数，

$$\boldsymbol{\delta}^u = \underset{\boldsymbol{\delta}^u \in \mathcal{B}(\epsilon)}{\operatorname{argmax}} -\mathcal{L}_{adv}\left(\mathbf{X}_h(t), \bigcup_{i=0}^{h} \boldsymbol{\delta}^u\right) - \lambda \times \gamma_v \qquad (28)$$

其中我们设定 $\lambda = 10^{10}$。与前文相同，式（28）根据攻击优化算法进行迭代求解。

数字与物理设置： 除了上述威胁模型外，我们还考虑两种不同的攻击设置——数字攻击和物理攻击。数字攻击设置（D-WB）描述了一种典型的白盒场景，其中假设攻击者能够入侵推理流水线，从而直接扰动模型的输入。相反，物理攻击设置（P-WB）描述了一种更现实的白盒场景：攻击者了解模型的架构和学习得到的权重，但无法直接扰动模型的输入。因此，攻击者必须在特定网格点处实际放置一定数量的设备，即对抗设备，以实现对抗扰动。P-WB 限制攻击者只能进行物理扰动，因此比 D-WB 更实用。尽管如此，由于其在文献中广受欢迎，

B. 对抗攻击

对于 D-WB 设置，我们在训练好的模型上评估了三种标准对抗攻击——FGSM、迭代快速梯度符号法和 PGD攻击。FGSM攻击简单、快速，并能生成可迁移的对抗扰动，因此非常适合用于实际的对抗评估。另一方面，PGD 是文献中针对未混淆模型（例如我们评估中使用的模型）提出的最强对抗攻击之一。对于 P-WB 设置，我们在不同模型架构上将上述攻击与我们新提出的 计算机视觉与模式识别会议 攻击进行比较。

C. 评估指标

测试损失： 为了评估模型 $\mathcal{F}$ 在测试数据 $\mathcal{D}_{test}$ 上的表现，我们使用一种常用指标——均方误差（MSE）来衡量模型输出与真实标签 $\mathbf{x}^{t+1}$ 之间的距离，定义如下：

$$\mathcal{L}(\mathcal{D}_{test}) = \frac{1}{|\mathcal{D}_{test}|} \sum_{\forall \mathbf{x}^t \in \mathcal{D}_{test}} (\mathcal{F}(\mathbf{X}_h(t)) - \mathbf{x}^{t+1})^2 \qquad (29)$$

$\mathcal{L}(\mathcal{D}_{test})$ 的值越小，说明模型学习得越好。

对抗损失： 对于对抗评估，我们令$\mathcal{D}^*_{test}$ 表示经对抗扰动的测试数据，并计算对抗 MSE，记为 $\mathcal{L}^*(\mathcal{D}^*_{test})$，以此衡量模型的鲁棒性。

$$\mathcal{L}^*(\mathcal{D}^*_{test}) = \frac{1}{|\mathcal{D}^*_{test}|} \sum_{\forall \mathbf{x}^t \in \mathcal{D}^*_{test}} (\mathcal{F}(\mathbf{X}_h(t)) - \mathbf{y}^{t+1}_{target})^2 \qquad (30)$$

其中，$\mathbf{y}^{t+1}_{target}$ 表示攻击者期望的目标输出。 $\mathcal{L}^*(\mathcal{D}^*_{test})$的值越小，表明模型对对抗扰动的鲁棒性 越强，反之亦然。

误接受率（FAR）： 为评估*CaV-detector* 捕获对抗性输入的效果，我们采用一种广泛使用的指标——误接受率（FAR）。其定义为：被检测器标记为未受扰动的对抗性输入数量占攻击者生成的对抗性输入总数的比例。此外，我们还使用 FAR 来量化对抗攻击逃避*CaV-detector*检测的效果。

D. 超参数

在本小节中，我们报告用于分析模型在标准场景和对抗场景下性能的关键超参数。

数据：在准备数据集时，我们使用的历史长度为 $h \in \{2, 5, 10, 15, 20\}$。

模型：我们基于多层感知机和STResnet架构训练不同的模型，并改变各模型的隐藏层数量。对于多层感知机模型，我们使用的隐藏层数量为 {3, 5, 10}。下文将具有 h个隐藏层的多层感知机模型记作*mlp-h*。对于STResnet模型，我们使用的隐藏残差块数量为 {1, 2, 3}，并分别将它们记作*STResnet-1,STResnet-2* 和*STResnet-3*。对于时空图卷积网络模型，我们试验了 {1, 3, 5, 10} 种不同的隐藏消息维度（参见 [12]对时空图卷积网络模型中隐藏消息的定义），并研究邻居数量 $d_A \in \{1,3、5、 10\}$ 的变化对时空图卷积网络模型准确率的影响，其中 d_A 表示时空图卷积网络模型假定能够相互通信的相邻节点数量。为便于后文引用，我们将隐藏消息维度为 m 、节点连通性为 d_A 的时空图卷积网络模型记作*tgcn-*(m, d_A)。

攻击： 对于对抗评估，我们针对不同的最大扰动进行实验，包括 $\epsilon \in \{0.01、0.03、0.05、0.07、0.1\}$（用于数字攻击设置），以及最大对抗设备预算 $b_d \in \{1000, 3000, 5000, 7000,10000\}$ （用于物理攻击设置）。我们还分析了改变最大攻击迭代次数 $N \in \{100, 250, 500, 750, 1000\}$ 对上述模型性能的影响。为便于后续引用，我们将

特定的攻击设置表示为 Attack-(ϵ, N)。例如，PGD- (0.1,1000) 表示一个PGD攻击者，其允许的最大扰动为0.1，迭代预算为1000。

V. 结果

我们首先通过报告原始/未扰动输入上的均方损失来建立基线。然后，我们评估这些模型在标准对抗攻击和新提出的计算机视觉与模式识别会议 攻击下的表现。最后，通过比较各自达到对抗目标所需的对抗设备数量，展示计算机视觉与模式识别会议 攻击相较于标准对抗攻击的有效性。

A. 预测模型性能

图8(a、b和c)比较了在具有不同历史长度的TaxiBJ-16数据集上训练的多层感知机、时空图卷积网络和时空残差网络模型（复杂度各不相同）在未扰动测试输入上的 $\mathcal{L}(\mathcal{D}_{test})$。对于本实验所选的超参数，我们未观察到模型复杂度与其在未扰动测试输入上的性能之间存在明确关系。我们还注意到，时空残差网络模型通常优于多层感知机和时空图卷积网络模型，这可归因于其架构中编码的先验使其能够捕获数据中的时空关系。尽管时空图卷积网络模型也能捕获数据中的时空模式，但与时空残差网络模型相比，其参数要少得多。

在图8中，我们观察到，随着输入历史长度增加，模型的 $\mathcal{L}(\mathcal{D}_{test})$通常也会增加，不过增幅很小。本文考察的三种模型架构均呈现这一趋势，偶尔也有例外。我们推测，历史长度增加会向模型提供更多输入信息，这可能导致训练过程中出现相互矛盾的梯度更新，使最终模型欠拟合。对于本就容易欠拟合的相对简单模型，随着 h 增加， $\mathcal{L}(\mathcal{D}_{test})$的增幅更为显著，这进一步验证了我们的假设。

B. D-WB 盲对抗攻击

在此实验中，我们假设攻击者不了解*CaV-detect* 机制，因此攻击时会假定模型是基础模型。图9(a-c)总结了我们对不同架构的人群流量预测模型——MLP-5、TGCN-(5,5) 和 STResnet-2——在不同扰动预算 $\epsilon \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$下进行四种对抗攻击的结果：FGSM-($\epsilon$,1)、i-FGSM-($\epsilon$, 500)、PGD-($\epsilon$, 500)以及计算机视觉与模式识别会议 attack-(ϵ,500)。总体而言，与 $\epsilon = 0$相比，深度人群流预测模型和其他深度学习模型一样，很容易受到对抗攻击；对于 $\epsilon > 0$，其$\mathcal{L}^*(\mathcal{D}^*_{test})$值显著更小，这说明了这一点。

如图所示，增大 ϵ 会使攻击更强，这一点可从 $\mathcal{L}^*(\mathcal{D}^*_{test})$的值相应降低看出。这些结果符合我们的直觉，因为更大的 ϵ 使攻击者能够更大程度地影响模型的输入，而模型的输入又决定其输出。总体而言，PGD攻击似乎是所有攻击中最强的，而我们的计算机视觉与模式识别会议 攻击似乎最弱。不过，正如我们稍后将看到的，三种标准攻击生成的对抗扰动都能被我们的*CaV-detect* 机制100%检测到，而计算机视觉与模式识别会议攻击生成的扰动则不会被检测到。

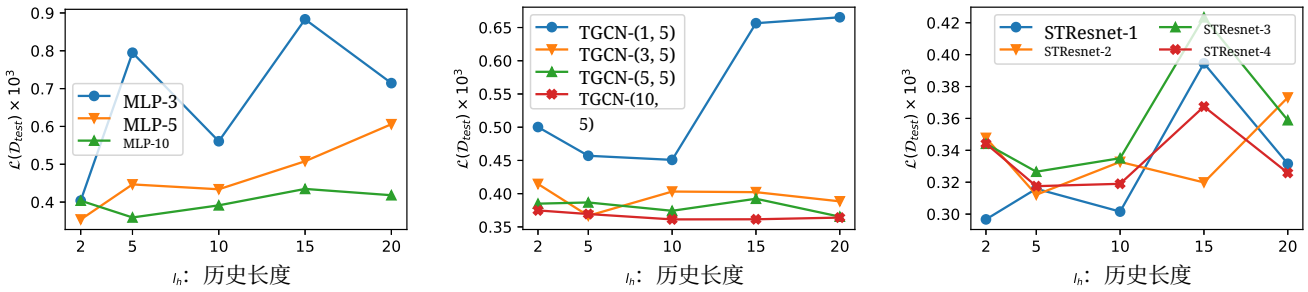
我们注意到，与 MLP-5 和 TGCN-(5,5) 相比，STResnet-2 模型的 $\mathcal{L}^*(\mathcal{D}^*_{test})$ 值较小。这表明 STResnet-2 模型相对更容易受到对抗扰动的影响，这似乎令人意外，因为我们观察到，与其他架构相比，STResnet-2 模型在未扰动数据集 $\mathcal{D}_{test}$ 上的表现相对更好。这些观察结果表明，人流预测模型可能存在准确率-鲁棒性权衡，这一现象在其他深度学习模型中也很常见 [5], [61]。

与其他架构不同，对于 MLP-5 模型，随着最大扰动预算 ϵ 增大，对抗损失 $\mathcal{L}^*(\mathcal{D}^*_{test})$ 略有上升。这是因为快速梯度符号法是单步攻击，攻击者计算出的梯度只能在有限的输入扰动范围内估计损失曲面。更大的扰动会使这些梯度变得不准确，从而降低攻击的效力。另一方面，迭代快速梯度符号法攻击者会在每次小幅扰动后迭代计算这些梯度，从而显著增强攻击效果。

检测 D-WB 盲攻击对抗扰动：在本实验中，我们评估一致性和有效性属性检测对抗扰动的效果。我们假设攻击者为图7所示的 D-WB盲攻击者。我们使用新颖的 *CaV-detect* 机制处理标准对抗攻击生成的对抗扰动，并报告 MLP-5、TGCN-(5,5) 和 STResnet- 2 的结果。总之，当错误拒绝率（FRR）设为≤0.5% 时，我们的 *CaV-detect* 机制对标准对抗攻击的误报率（FAR）为 0%，对我们提出的 计算机视觉与模式识别会议攻击的误报率为 >99.7%。

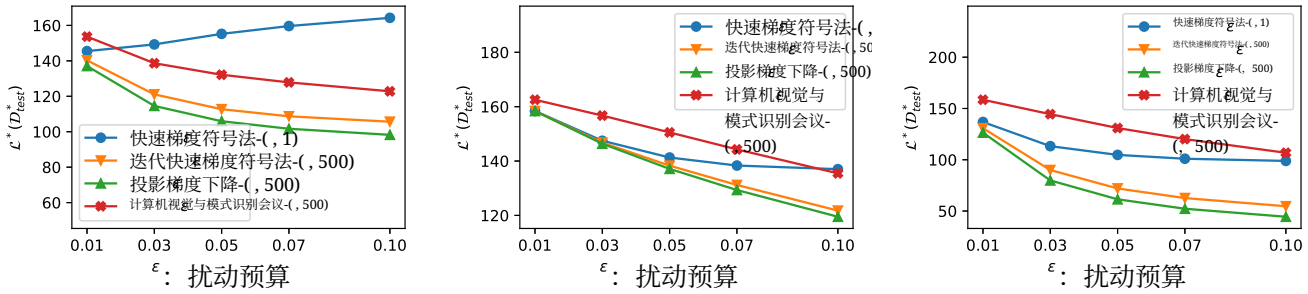
在图 10 中，我们仅报告有效性检查机制的误报率。这是因为在我们的实验（D-WB盲攻击者）中，一致性检查机制对标准对抗攻击的误报率始终为 0%，而对计算机视觉与模式识别会议攻击的误报率始终为 100%。因此，*CaV-detect* 对标准对抗攻击的总体误报率为 0%，而对计算机视觉与模式识别会议 攻击的总体误报率等于有效性检查机制的误报率。

在图 10 中，我们观察到，随着 ϵ 增加，对抗扰动会变得越来越无效。这并不令人意外，因为人群流动输入 $\mathbf{X}_h(t)$ 初始时是有效的，而引入幅度更大的无效扰动，会对扰动输入的有效性产生更显著的影响。对于 MLP-5 模型，我们观察到，与 TGCN-(5,5) 和 STResnet-2 模型相比，其生成的对抗扰动相对更有效。我们推测，由于架构相对简单，MLP-5 模型学习到的特征大多是线性的，因此关于模型输入的梯度也是线性的。根据式 (13) 的定义，如果$\mathbf{X}_h(t)$ 有效，那么它的线性倍数也有效。



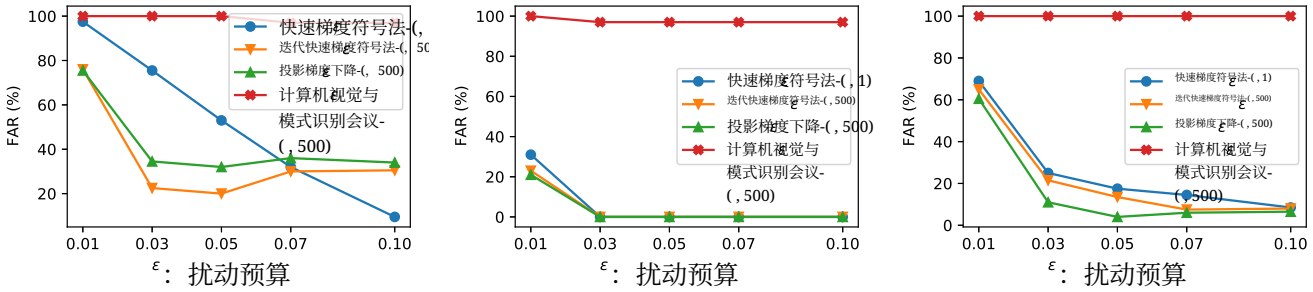
(a) 多层感知机（MLP） (b) 时间图卷积网络 (c) 时空 ResNet（STResnet）

图8：对**模型损失**，$\mathcal{L}(\mathcal{D}_{test})$（公式-(29)），在**原始/未扰动测试集**，$\mathcal{D}_{test}$，对于不同模型复杂度，随着预定义历史长度 h的增加。（设置：数据集为 TaxiBJ-16。）在 $\mathcal{D}_{test}$ 上未观察到模型复杂度与其性能之间存在严格关系。输入历史长度增加时，$\mathcal{L}(\mathcal{D}_{test})$略有增加，表明模型性能有所下降。在这三种架构中，*STResnet* 的表现最佳。



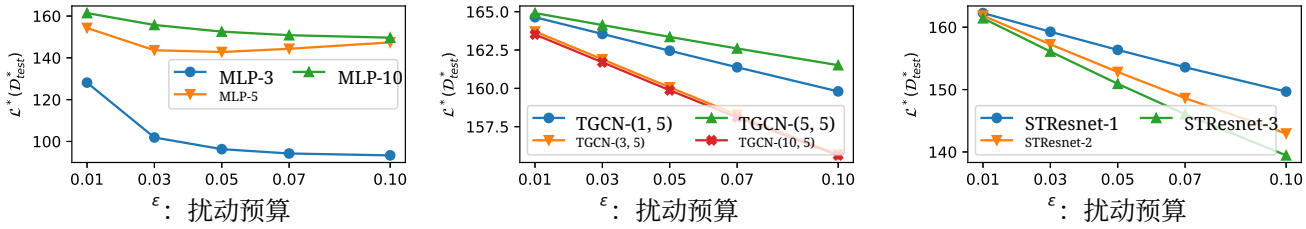
(a) MLP-5模型 (b) TGCN-(5,5)模型 (c) STResnet-2模型

图9：比较**对抗损失**、$\mathcal{L}^*(\mathcal{D}^*_{test})$（公式-(30)）在**扰动数据集**上 $\mathcal{D}^*_{test}$，由针对不同模型架构的不同攻击所产生，并随着 ϵ 增加而变化，假设**D-WB盲攻击者**。（设置：数据集为 TaxiBJ-16；h 为 5。）深度人群流预测模型容易受到对抗攻击。增大 ϵ 会使攻击更强。在所考虑的架构中，*TGCN-(5,5)*模型最为稳健。

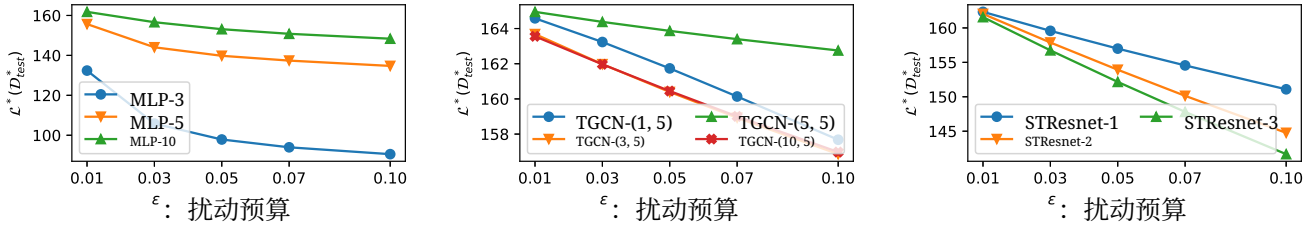


（a）MLP-5模型 （b）TGCN-(5,5)模型 （c）STResnet-2模型

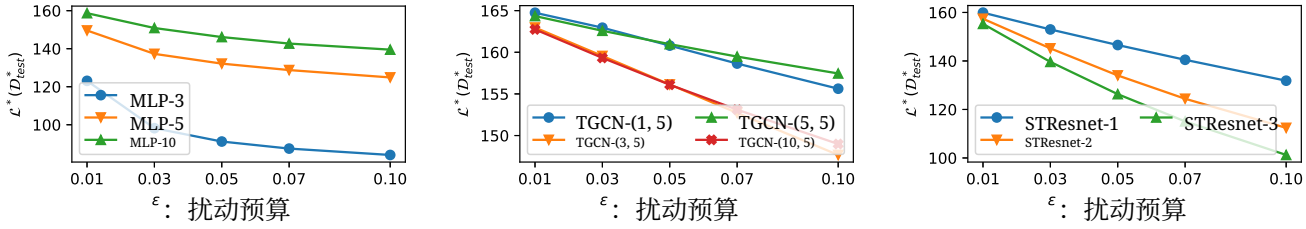
图 10：CaV-detect机制针对扰动输入的 $\mathcal{D}^*_{test}$，由**D-WB盲攻击者**生成。（设置：数据集为 TaxiBJ-16。h 为 5）。随着 ϵ 增加，对抗扰动变得越来越无效。一致性检查机制的误报率始终为 *0%*，因此我们只报告有效性检查机制的误报率。



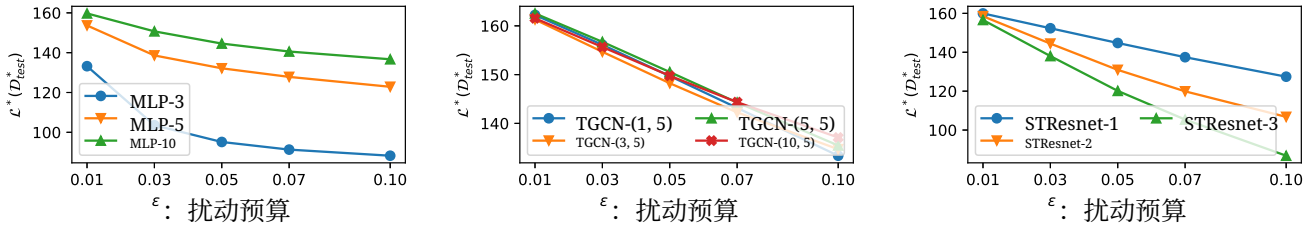
（a）使用 FGSM-(ϵ, 1) 攻击，对不同模型架构下扰动数据集 $\mathcal{D}^*_{test}$ 上的对抗损失 $\mathcal{L}^*(\mathcal{D}^*_{test})$ 进行比较。



（b）使用 iFGSM-(ϵ, 500) 攻击，对不同模型架构下扰动数据集 $\mathcal{D}^*_{test}$ 上的对抗损失 $\mathcal{L}^*(\mathcal{D}^*_{test})$ 进行比较。



（c）对于不同的模型架构，使用 PGD-(ϵ, 500) 攻击，在扰动数据集 $\mathcal{D}^*_{test}$ 上比较对抗损失 $\mathcal{L}^*(\mathcal{D}^*_{test})$。



（d）比较对抗损失 $\mathcal{L}^*(\mathcal{D}^*_{test})$，对于不同的模型架构，使用 CVPR-($\epsilon$, 500) 攻击，在扰动数据集 $\mathcal{D}^*_{test}$ 上 .

图11：比较**对抗损失**，$\mathcal{L}^*(\mathcal{D}^*_{test})$，在假设**D-WB盲攻击者**的情况下，$\mathcal{D}^*_{test}$，通过不同的攻击算法（a）—（d）比较不同模型架构，随着 ϵ 增加，并假设**D-WB自适应攻击者**。（设置：数据集为 TaxiBJ-16；l_h 为 5）。论文中假设的三种架构中，*TGCN-(5,5)* 模型对不同攻击表现出最强的自适应对抗鲁棒性，其次是 *STResnet-2* 模型。。

*C. D-WB*自适应对抗攻击

在本实验中，我们假设攻击者知晓*CaV-detect* 机制，并能够自适应地重新制定攻击方法，从而独自欺骗所有人群流量预测模型和*CaV-detect* 机制。图9(a-d)总结了我们在本文假设的人群流量预测模型上，针对不同扰动预算 $\epsilon \in \{0.01, 0.03, 0.05, 0.07, 0.1\}$进行四种不同对抗攻击的结果——FGSM-($\epsilon$,1)、i-FGSM-($\epsilon$,500)、PGD-($\epsilon$, 500)和计算机视觉与模式识别会议 攻击-($\epsilon$,500)。

如前所述，增加 ϵ 会增强攻击的力度，从而更显著地影响模型的输出，这一点可从 $\mathcal{L}^*(\mathcal{D}^*_{test})$ 值降低看出。与我们此前的观察（STResnet- 2模型的鲁棒性最差）相反，我们注意到，面对自适应攻击时，STResnet模型的鲁棒性与多层感知机模型不相上下，甚至更好。在论文假设的三种架构中，时空图卷积网络模型表现出最强的自适应对抗鲁棒性。

对于 MLP 架构，我们注意到，与 MLP-3 和 MLP-5 模型相比，更深的 MLP-10 模型具有更强的鲁棒性。这些观察结果与先前的一项研究 [5],一致，该研究表明，提高模型复杂度会略微增强模型的鲁棒性。我们发现，本文所考虑的所有标准攻击都具有相近的对抗性能，这一结果有些反直觉，因为 i-FGSM 和 PGD 属于迭代攻击，通常被认为比 FGSM攻击更强。我们将此归因于 MLP架构本身较为简单，无法从输入数据中学习到鲁棒特征，因此相对简单的攻击也能同样有效地操纵它。

对于 TGCN 架构，我们没有观察到模型复杂度增加会对其对抗性鲁棒性产生明确而一致的影响。例如，TGCN-(5,5)传递 5 维隐藏消息，与较简单的 TGCN-(1,5) 和较复杂的 TGCN-(10,5) 相比，通常具有更强的鲁棒性；后两者分别传递 1 维和 10 维隐藏消息。总体而言，我们提出的 计算机视觉与模式识别会议攻击显著优于自适应 PGD攻击，而后者又大幅优于自适应 i-FGSM攻击。i-FGSM攻击和 FGSM攻击的表现相近，在某些情况下，i-FGSM略胜于 FGSM。

对于 STResnet架构，STResnet-1表现出显著更强的对抗鲁棒性，其次是 STResnet-2，而 STResnet-3 的鲁棒性则明显较弱，这可归因于潜在DNN层的对抗脆弱性增加 [62]。我们再次观察到，自适应投影梯度下降和计算机视觉与模式识别会议攻击的效果显著优于自适应快速梯度符号法和自适应迭代快速梯度符号法攻击，尤其是 $\epsilon = 0.1$。例如，对于 STResnet-2 模型，自适应投影梯度下降和计算机视觉与模式识别会议 攻击的 $\mathcal{L}^*(\mathcal{D}^*_{test}) \approx 100$，而自适应快速梯度符号法和自适应迭代快速梯度符号法攻击的 $\max \mathcal{L}^*(\mathcal{D}^*_{test}) \approx 145$。

检测D-WB自适应对抗扰动：在本实验中，我们评估能否依据先前形式化定义的一致性和有效性属性，检测经过自适应扰动的对抗性输入。我们假设攻击者采用图7所示的D-WB自适应攻击。我们将自适应对抗攻击生成的对抗扰动输入 *CaV-detect* 进行处理，并报告 MLP-5、TGCN-(5,5) 和 STResnet-2 模型的结果。总之，当错误拒绝率（FRR）设为 $\leq$0.5%时，我们的*CaV-detect*机制针对自适应对抗攻击表现出显著更高的误报率（≈80%-100%），相较于针对非自适应标准攻击的误报率（≈0%），且针对我们提出的计算机视觉与模式识别会议攻击的误报率为 >99.7%。

在图 12 中，我们仅报告自适应攻击设置下有效性检查机制的误报率。这是因为在我们的实验中（D-WB 自适应攻击者），由于通用对抗扰动，一致性检查机制面对所有自适应对抗攻击时的误报率始终为 100%。因此，*CaV-detect* 的整体误报率等于有效性检查机制的误报率。

如前所见，增加 ϵ 会显著降低误报率，表明对抗扰动变得越来越无效。与针对*CaV-Detect*的盲攻击不同，D-WB 自适应攻击生成的对抗性输入能够躲避检测机制；对于标准对抗攻击，其误报率约为 80%。我们特别将这一点归因于对标准攻击新提出的自适应改进——将对抗扰动通用化以及拉格朗日优化。

然而，尽管采用了D-WB自适应算法，我们注意到，FGSM攻击对*CaV-detect* 的效果并不理想，如误报率显著降低所示；尤其是在TGCN-(5,5)和STResnet-2模型上，当 $\epsilon = 0.1$时，误报率降至0%。同样，我们将其归因于FGSM是一次性攻击，导致大幅扰动下的梯度不精确，从而显著降低了攻击的有效性。

VI. 讨论

A. 人群流量预测可视化

图 13 对比了不同人流预测模型在不同未来时刻的预测流入状态与实际流入状态（记录于未来的真实值），涉及原始输入以及由不同自适应攻击生成的扰动输入；这些攻击旨在尽可能提高预测流入状态，同时使扰动保持 $\boldsymbol{\delta} \in \mathcal{B}_\infty(\epsilon)$。定性分析表明，计算机视觉与模式识别会议攻击优于其他攻击，因为无论模型架构如何，经计算机视觉与模式识别会议攻击的输入通常都呈现最高的预测流入状态。此外，与另外两个模型相比，TGCN-(5,5) 模型对于自适应攻击生成的一致且有效的扰动更加鲁棒。另外，我们注意到，当原始预测流入状态相对较小时，预测流入状态受计算机视觉与模式识别会议 扰动的影响更大。

有趣的是，我们注意到，对于本实验中考察的所有模型，对抗扰动输入所产生的预测流入状态通常与原始预测流入状态高度相关，相关性可能为正（图 13(b,c)），也可能为负（图 13(a) 中 MLP-5 上的 FGSM-(0.1,1) 和 CVPR-(0,1,500)）。基于这些观察，我们推测人群流量预测模型的表达能力有限——无论输入如何，模型都无法生成某些输出。

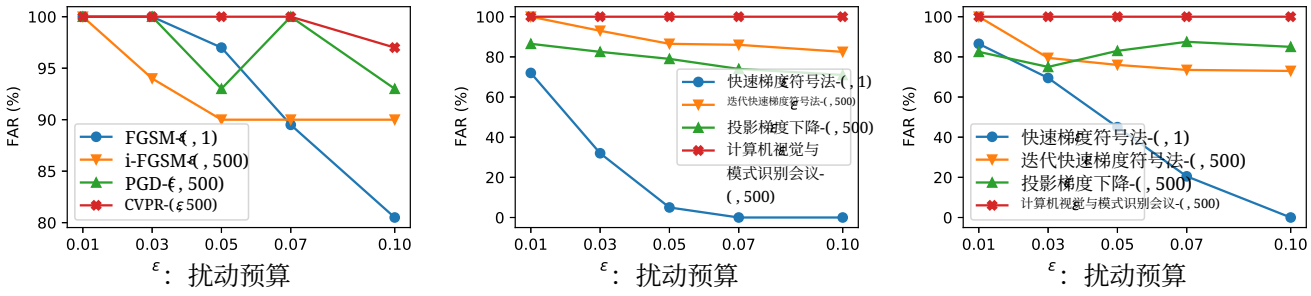
B. 历史长度 h 对对抗损失的影响

我们分析改变历史长度 h 对不同攻击的对抗损失的影响。我们针对不同的历史长度 $h \in \{2, 5, 10, 15, 20\}$ 训练了三个不同的模型——MLP-5、TGCN-(5,5) 和 STResnet-2——并使用本文所考察的对抗攻击对这些模型进行攻击。请注意，对于每个 h，我们都会训练一个新模型，因为针对不同 h 值训练的模型，其输入彼此不兼容。在本实验中，我们为所有攻击设置 ϵ=0.1 和 N=500 。图 14(a-c) 分别报告了上述三个模型上各攻击的 $\mathcal{L}^*(\mathcal{D}^*_{test})$ 值。

我们发现，模型的对抗鲁棒性与其训练时采用的历史长度之间不存在严格的关系。不过，我们注意到，通常增加模型的历史长度可能会使其对对抗扰动略微更具鲁棒性。乍看之下，这似乎有悖直觉，因为更长的历史长度允许攻击者向输入中添加更多扰动。然而，回想一下我们在第 V-A 节中的观察，我们将鲁棒性的提升归因于历史长度增加时模型性能的下降，而这反过来可能是由准确率-鲁棒性权衡 [5], [61] 导致的。

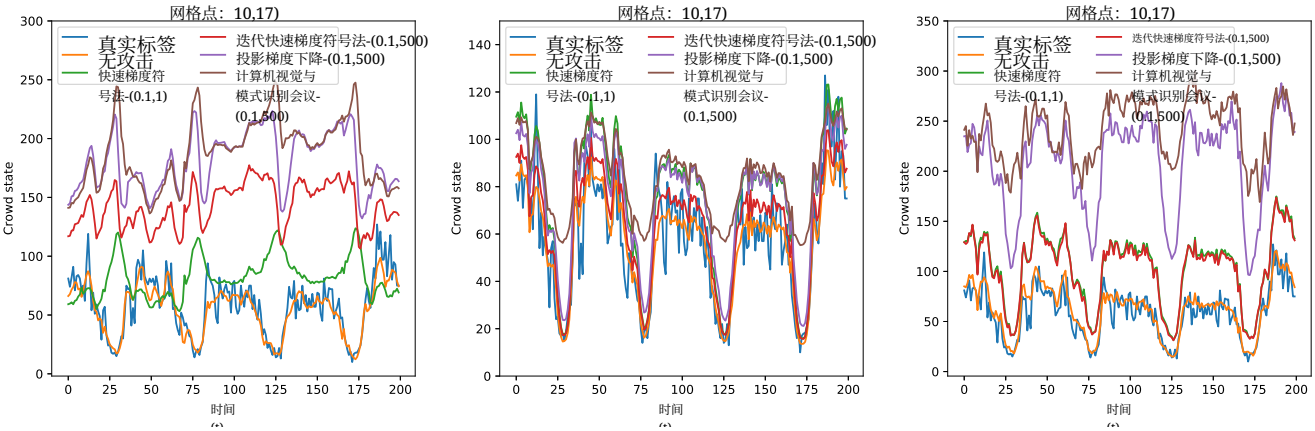
C. 对抗攻击速度

图 15(a-c) 比较了不同攻击优化对抗损失的速度，以及迭代次数的影响



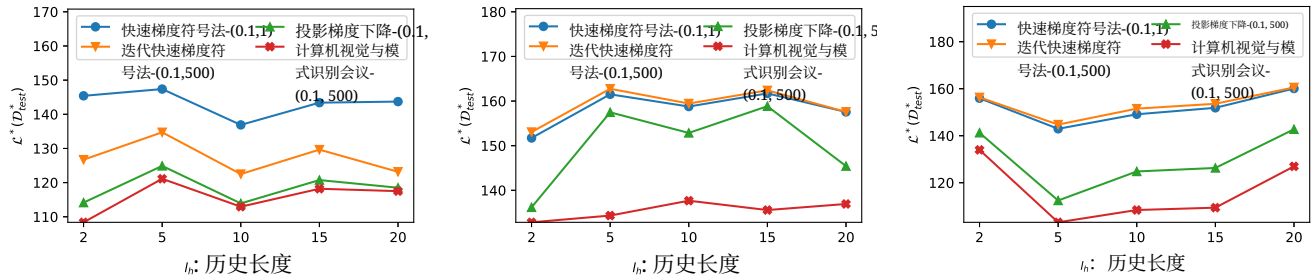
(a) MLP-5　(b) TGCN-(5,5)　(c) STResnet-2

图 12：CaV-detect 的误接受率（FAR）机制针对扰动输入的表现，$\mathcal{D}^*_{test}$，由**D-WB 自适应攻击者**生成。（设置：数据集为 TaxiBJ-16。 h为 5。）随着 ϵ 增大，对抗扰动变得越来越无效。一致性检查机制的 *FAR* 始终为 *100%*，因此我们仅报告有效性检查机制的 *FAR*。



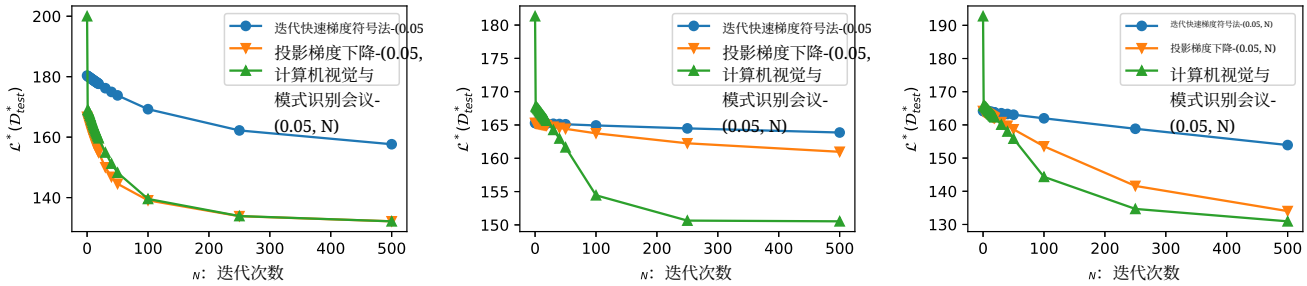
(a) MLP-5模型　(b) TGCN-(5,5)模型　(c) STResnet-2模型

图 13：将不同架构的人群流量预测模型预测的流入状态与实际流入状态（未来记录的流入状态）进行可视化对比。“无攻击”表示在假设存在**D-WB自适应攻击者**的情况下，原始/未扰动输入对应的预测流入状态。（设置：数据集为 TaxiBJ-16； h 为 5； ϵ 为 0.1）。计算机视觉与模式识别会议攻击优于其他攻击。与另外两个模型相比，*TGCN-(5,5)*模型对一致且有效的对抗攻击更加鲁棒。



(a) MLP-5模型　(b) TGCN-(5,5)模型　(c) STResnet-2模型

图14：比较**对抗损失**， $\mathcal{L}^*(\mathcal{D}^*_{test})$ ，在**由不同攻击针对采用不同历史长度训练的不同模型（具有不同架构）生成的扰动测试输入**上， h。（设置：数据集为 TaxiBJ-16； ϵ 为 0.1）。通常，随着输入历史长度增加， $\mathcal{L}^*(\mathcal{D}^*_{test})$ 会略微增加，这表明采用更长历史长度训练的模型对对抗扰动略微更加鲁棒。



(a) MLP-5模型　(b) TGCN-(5,5)模型　(c) STResnet-2模型

图15：假设一**名 D-WB自适应攻击者**，随着攻击的进行，比较针对不同模型架构的不同攻击算法在扰动数据集 $\mathcal{D}^*_{test}$上导致对抗损失 $\mathcal{L}^*(\mathcal{D}^*_{test})$下降的情况。（设置：数据集为TaxiBJ-16；$h$ 为5； ϵ 为0.05）。在给定扰动预算的情况下，计算机视觉与模式识别会议攻击在对抗损失和速度方面始终优于其他攻击。

分别针对三种架构各异的模型——MLP-5、TGCN-(5,5)和STResnet-2。随着攻击的进行，无论采用哪种攻击，生成的对抗扰动都会随着每次迭代而有所改善，这一点可从对抗损失的下降看出，符合预期。计算机视觉与模式识别会议 攻击始终以显著优势胜过另外两种攻击，这一点在TGCN-(5,5)和STResnet-2模型上尤为明显。我们认为，这是因为计算机视觉与模式识别会议 攻击在扰动生成机制中明确引入了人群流有效性先验，从而生成一致且有效的对抗扰动，使优化器不受式(27)中的一致性和无效性分数所限制。

D. 对抗攻击的物理可实现性

为进一步了解对抗攻击的严重性，在本实验中，我们研究了严格受限的威胁模型下对抗攻击的有效性。该模型假设物理攻击者可以读取模型权重和输入，但无法扰动模型的输入。图7中的 P-WB-自适应攻击展示了这种威胁模型。此外，我们进一步限制攻击者，假设其处于有限查询设置 [7]，最多只能查询模型 20 次。我们还假设攻击者的设备预算有限， b_d，即定义攻击者能够实际控制的设备数量，我们将这些设备称为对抗设备。攻击者的目标是通过实际移动对抗设备（以模拟对抗扰动）来欺骗人群预测模型。我们改变 $b_d \in \{500, 1000, 5000, 10000, 15000\}$，并在图16中报告两种攻击——投影梯度下降-($b_d$, 20) 和 CVPR-($b_d$, 20)——在不同 b_d 值下的对抗损失。

有趣的是，我们注意到，对于相对较小的设备预算 b_d，在本实验考虑的三种不同架构的模型中，PGD攻击在 $\mathcal{L}^*(\mathcal{D}^*_{test})$值方面始终优于计算机视觉与模式识别会议 攻击。我们将此归因于两个原因。首先，与计算机视觉与模式识别会议 攻击生成的对抗扰动相比，PGD攻击生成的对抗扰动无效性相对更高；图16最后一行展示了这一点，其中报告了以下内容的误报率：

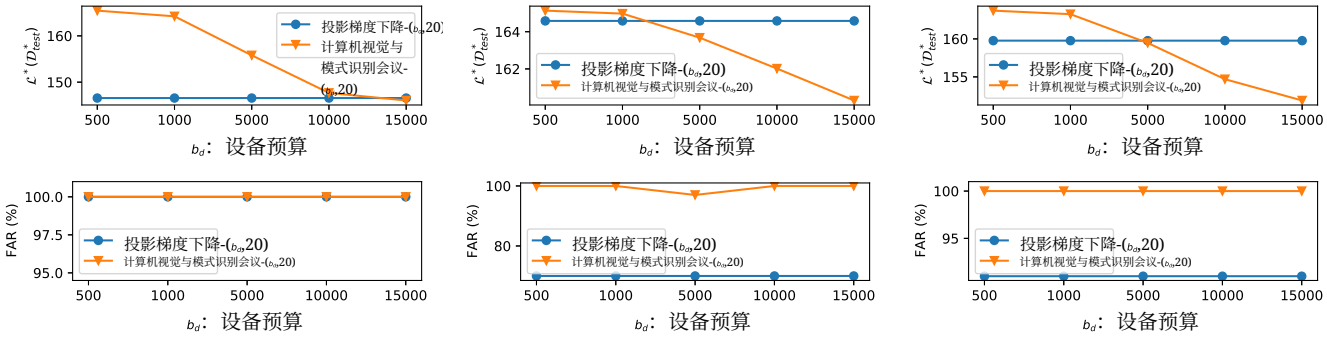
CaV-detect 机制。其次，式（21）中提出的流出扰动生成机制隐式地对 $\boldsymbol{\delta}_{in}$ 和 $\boldsymbol{\delta}_{out}$施加了额外约束。尽管这些额外约束从长远来看是有益的，尤其是对于更强的对手而言（如前文图15所示），但当扰动预算过于有限时，它们可能会损害攻击效果。

与图11中的 $\mathcal{L}^*(\mathcal{D}^*_{test})$数值相比，图16中的 $\mathcal{L}^*(\mathcal{D}^*_{test})$明显更小，这主要是因为攻击者的查询预算和设备预算有限。这表明，尽管在人群流动预测模型在物理场景下易受一致且有效的对抗扰动影响，但与数字攻击设置相比，实现目标输出要困难得多。

E. 人流预测模型的有限表达能力

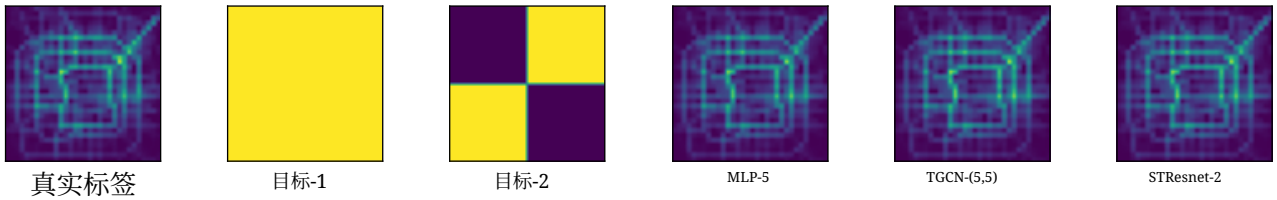
在本实验中，我们展示了三种人流预测模型——MLP-5、TGCN-(5,5) 和 STResnet-2——的表达能力有限。我们将表达能力定义为模型在给定适当输入时生成特定输出的能力。图17(b)显示，三种模型给出的预测都接近未来记录的真实输出（图17(a)）。为展示这些模型表达能力有限，我们假设存在一个强大的 PGD-(1,500)对手，拥有$\epsilon = 1$，因此该对手可以在*CaV-detect*盲威胁模型下任意更改输入；这样一来，攻击者只需优化输入，而无需考虑*CaV-detect*机制。此外，我们为攻击者设定了两个对抗目标状态——图17(a)中的目标-1和目标-2——让其尽可能扰动输入，从而使模型输出目标状态。图17(c,d)报告了上述模型在接收由针对目标-1和目标-2的 PGD-(1,500)攻击分别生成的对抗扰动输入后的输出预测。

理想情况下，人们可能会认为，由于对手可以完全控制输入，因此对手能够操纵模型，使其产生任意所需的输出。然而，图17(c,d)所报告的结果表明，人流预测模型并非如此。例如，输出

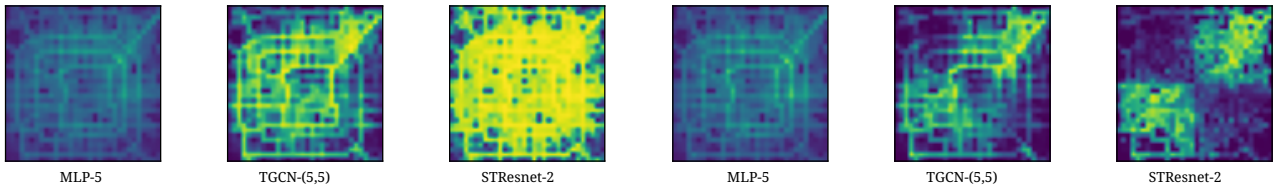


(a) MLP-5模型　(b) TGCN-(5,5)模型　(c) STResnet-2模型

图16：在不同设备预算下，比较针对不同模型架构的PGD攻击与计算机视觉与模式识别会议 攻击的物理合理性，比较指标为对抗损失、 $\mathcal{L}^*(\mathcal{D}^*_{test})$以及假设**D-WB自适应攻击者**时*CaV-detect* 的误报率（FAR）。（设置：数据集为TaxiBJ-16； h 为5； b_d 为攻击者在物理上最多可控制的设备数量。）



(a) 展示真实流入状态（记录于未来）以及对抗目标流入状态。(b) 比较三个模型——MLP-5、TGCN-(5,5)和STResnet-2——基于原始输入预测的流入状态。



(c) 比较三个模型基于针对目标-1优化的PGD盲攻击输入预测的流入状态。(d) 比较三个模型基于针对目标-2优化的PGD盲攻击输入预测的流入状态。

图17：说明不同架构的人群流量预测模型的有限表达能力（即在对输入拥有无限控制能力的情况下，模型产生所需输出的能力）。（设置：数据集为TaxiBJ-16； h 为5；攻击为PGD-(1,500)； ϵ 为1——表示对输入拥有无限控制能力）。在本实验考察的三个模型中，*STResnet-2*模型的表达能力最强，而*MLP-5*模型的表达能力最弱。

MLP-5模型的输出与目标存在显著差异，这表明MLP-5模型无法生成目标输出。我们认为，MLP-5模型的表达能力非常有限。尽管与MLP-5模型相比，TGCN-(5,5)和STResnet-2模型的输出更接近目标输出，但它们仍然缺乏精确生成目标输出所需的表达能力。我们将此归因于TaxiBJ数据集中输出大多聚集且高度相似。总体而言，我们观察到，STResnet-2模型在三者中表达能力最强，这也解释了为什么STResnet模型相比TGCN模型在对抗性方面更加脆弱（如图11所示）。

有趣的是，对于两个目标状态，攻击者只能从 MLP-5模型获得相似的输出。我们将其归因于 MLP-5架构过于简单，无法捕捉数据集中有意义且具有普适性的人群流量模式来自数据集。

VII. 结论

本文研究了三种不同架构的人群流量预测模型的对抗脆弱性：多层感知机、时序图卷积神经网络和时空ResNet。我们深入分析了模型复杂度和人群流动数据历史长度的变化对所得模型性能和对抗鲁棒性的影响，并发现人群流量预测模型与其他深度学习模型一样，很容易受到对抗攻击。其次，我们识别并规范化了人流输入的两种新属性——一致性和有效性，可用于检测经过对抗扰动的输入。因此，我们提出了 *CaV-detect* ，它可以检测对抗性输入，误报率为

0%。通过分析输入的一致性和有效性，*CaV-detect* 可将误报率为 ∼80-100%的输入判定为对抗扰动输入——只有同时满足一致性和有效性的模型输入才被视为未受扰动。随后，我们对标准对抗攻击进行自适应修改，以规避 *CaV-detect* 。最后，通过在对抗扰动生成机制中编码一致性与有效性先验，我们提出了 计算机视觉与模式识别会议 攻击，这是一种一致、有效且物理可实现的对抗攻击，在目标对抗损失和 *CaV-detect*的误报率两方面均优于自适应标准攻击。最后，我们深入探讨了针对人群流动预测模型的对抗攻击，并表明人群流量预测模型的表达能力有限，且可通过模拟通用扰动在物理上实现。

致谢

本出版物得以完成，离不开卡塔尔国家研究基金（卡塔尔基金会成员）提供的 NPRP 项目资助，项目编号为 [13S- 0206-200273] 。本文所述观点仅由作者负责。

参考文献

[1] S. Saadatnejad、M. Bahari、P. Khorsandi、M. Saneian、S.-M. Moosavi- Dezfooli 和 A. Alahi，“具备社会意识的轨迹预测模型是否真的具备社会意识？”*Transportation Research Part C: EmergingTech- nologies*，第 141 卷，文章编号 103705，2022年。[2] J. Zhang、Y. Zheng 和 D. Qi，“用于全市人群流量预测的深度时空残差网络”，载于第三十一届 *AAAI* 人工智能会议，2017年。[3] W. Jiang，“Taxibj21：一个基于北京出租车 GPS 轨迹的开放人群流量数据集”，*Internet Technology Letters*，第 5 卷，297，2022 2，第 e .[4] N. A. Petrick、S. Akbar、K. H. Cha、S. Nofech-Mozes、B. Sahiner、M. A. Gavrielides、J. Kalpathy-Cramer、K. Drukker、A. L. Martel等，“SPIE-AAPM-NCI BreastPathQ 挑战赛：针对新辅助治疗后乳腺癌组织学图像中肿瘤细胞密度定量评估的图像分析挑战赛”，*Journal of Medical Imaging*，第 8 卷，第 3 期，文章编号 034501，2021年。[5] H. Ali、M. S. Khan、A. AlGhadhban、M. Alazmi、A. Alzamil、K. Al- Utaibi 和 J. Qadir，“你们的假新闻检测器都是我们的：在黑盒环境下评估假新闻检测器的对抗鲁棒性”，*IEEE Access*，第 9 卷，第 81 678–81 692 页，2021年。[6] H. Ali、M. S. Khan、A. Al-Fuqaha 和 J. Qadir，“Tamp-X：通过篡改的激活值攻击可解释自然语言分类器”，*Computers & Security*，文章编号 102791，2022年。[7] F. Khalid、H. Ali、M. A. Hanif、S. Rehman、R. Ahmed 和 M. Shafique，“FADEC：一种用于对抗性机器学习的快速基于决策的攻击”，载于*2020*年国际神经网络联合会议（*IJCNN*）。IEEE，2020年，第 1–8 页。[8] M. A. Butt、A. Qayyum、H. Ali、A. Al-Fuqaha 和 J. Qadir，“迈向安全、隐私可信且以人为中心的嵌入式机器学习：情绪感知人脸识别案例研究”，*Computers & Security*，第 125 卷，文章编号 103058，2023年。[9] Y. Li、R. Yu、C. Shahabi 和 Y. Liu，“扩散卷积循环神经网络：数据驱动的交通流预测”，载于第六届国际学习表征会议，*ICLR 2018*，温哥华，不列颠哥伦比亚省，加拿大，*2018*年*4*月*30*日至*5*月*3*日，会议论文集。OpenReview.net，2018年。 [在线]。网址：https://openreview.net/forum?id=SJiHXGWAZ [10] A. Ilyas、S. Santurkar、D. Tsipras、L. Engstrom、B. Tran 和 A. Madry，“对抗样本并非漏洞，而是特征”，神经信息处理系统进展，第 32 卷，2019年。[11] C. Szegedy、W. Zaremba、I. Sutskever、J. Bruna、D. Erhan、I. J. Goodfellow 和 R. Fergus，“神经网络的有趣特性”，载于第二届国际学习表征会议，*ICLR 2014*，班夫，*AB*，加拿大，*2014*年*4*月*14*日至*16*日，会议论文集，Y. Bengio 和 Y. LeCun 编，2014年。 [在线]。网址：http://arxiv.org/abs/1312.6199

[12] L. Zhao、Y. Song、C. Zhang、Y. Liu、P. Wang、T. Lin、M. Deng 和 H. Li，“T-GCN：用于交通预测的时间图卷积网络”，IEEE智能交通系统汇刊，第21卷，第9期，第3848–3858页，2019年。[13] L. Mourad、H. Qi、Y. Shen 和 B. Yin，“Astir：用于人流预测的时空数据挖掘”，IEEE Access，第7卷，第175 159–175 165页，2019年。[14] A. Athalye、N. Carlini 和 D. Wagner，“混淆梯度带来虚假的安全感：规避对抗样本防御”，载于《国际机器学习会议》。PMLR，2018年，第274–283页。[15] F. Tramer、N. Carlini、W. Brendel 和 A. Madry，“关于针对对抗样本防御的自适应攻击”，《神经信息处理系统进展》，第33卷，第1633–1645页，2020年。[16] S.-M. Moosavi-Dezfooli、A. Fawzi、O. Fawzi 和 P. Frossard，“通用对抗扰动”，载于《IEEE计算机视觉与模式识别会议论文集》，2017年，第1765–1773页。[17] M. Chen、X. Yu 和 Y. Liu，“PCNN：用于短期交通拥堵预测的深度卷积网络”，IEEE智能交通系统汇刊，第19卷，第11期，第3550–3559页，2018年。[18] E. Onieva、V. Milan´es、J. Villagra、J. P´erez 和 J. Godoy，“交叉路口车辆模糊决策系统的遗传优化”，Expert Systems with Applications，第39卷，第18期，第13 148–13 157页，2012年。[19] X. Zhang、E. Onieva、A. Perallos、E. Osaba 和 V. C. Lee，“采用遗传算法优化的用于短期交通拥堵预测的分层模糊规则系统”，Transportation Research Part C: Emerging Technologies，第43卷，第127–142页，2014年。[20] E. Onieva、P. Lopez-Garcia、A. Masegosa、E. Osaba 和 A. Perallos，“进化模糊分类方法与清晰规则分类方法在拥堵预测中性能的比较研究”，《工程学程序汇刊》，第14卷，第4458–4467页，2016年。[21] P. Lopez-Garcia、E. Onieva、E. Osaba、A. D. Masegosa 和 A. Perallos，“ 利用遗传算法和交叉熵进行短期交通拥堵预测的混合方法 ”，IEEE智能交通系统汇刊，第17卷，第2期，第557–569页，2015年。[22] Y. Zheng、Y. Li、C.-M. Own、Z. Meng 和 M. Gao，“基于平衡马尔可夫链的交通拥堵模型实时预测与导航”，International Journal of Distributed Sensor Networks，第14卷，第4期，文章编号1550147718769784，2018年。[23] J. F. Zaki、A. Ali-Eldin、S. E. Hussein、S. F. Saraya 和 F. F. Areed，“基于隐马尔可夫模型和对比度度量的交通拥堵预测”，Ain Shams Engineering Journal，第11卷，第3期，第535–551页，2020年。[24] S. Sun、J. Chen 和 J. Sun，“基于GPS轨迹数据的交通拥堵预测”，International Journal of Distributed Sensor Networks，第15卷，第5期，文章编号1550147719847440，2019年。[25] Y. Qi 和 S. Ishak，“高速公路交通状况短期预测的隐马尔可夫模型”，Transportation Research Part C: Emerging Technologies，第43卷，第95–111页，2014年。[26] S. Yang，“关于交通拥堵预测中的特征选择”，Transportation Research Part C: Emerging Technologies，第26卷，第160–169页，2013年。[27] L. Zhu、R. Krishnan、F. Guo、J. W. Polak 和 A. Sivakumar，“异构城市交通中周期性拥堵的早期识别”，载于2019年IEEE智能交通系统会议（ITSC）。IEEE，2019年，第4392–4397页。[28] S. Sun、C. Zhang 和 G. Yu，“交通流预测的贝叶斯网络方法”，IEEE智能交通系统汇刊，第7卷，第1期，第124–132页，2006年。[29] G. Asencio-Cort´es、E. Florido、A. Troncoso 和 F. Mart´ınez- ´Alvarez，“利用集成学习预测城市交通拥堵的新方法论”，Soft Computing，第20卷，第11期，第4205–4216页，2016年。[30] J. Kim 和 G. Wang，“利用贝叶斯网络诊断和预测城市道路网络中的交通拥堵”，Transportation Research Record，第2595卷，第1期，第108–118页，2016年。[31] X. Yang、S. Luo、K. Gao、T. Qiao 和 X. Chen，“数据科学技术在交通拥堵智能预测中的应用”，Journal of Advanced Transportation，第2019卷，2019年。[32] K. M. Nadeem 和 T. P. Fowdur，“实时自适应交通拥堵预测算法的性能分析”，Journal of Information and Communication Technology，第17卷，第3期，第493–511页，2018年。[33] J. Lee、B. Hong、K. Lee 和 Y.-J. Jang，“利用天气数据的交通拥堵预测模型”，载于2015年IEEE数据科学与数据密集型系统国际会议。IEEE，2015年，第81–88页。[34] P. Zhang 和 Z. S. Qian，“以用户为中心的相互依赖城市系统：利用分时电力使用数据预测早晨道路

拥堵”，Transportation Research Part C: Emerging Technologies，第92卷，第392–411页，2018年。[35] S. Jain、S. S. Jain 和 G. Jain，“基于起点和终点的交通拥堵建模”，《工程学程序汇刊》，第187卷，450，2017年442页。–。[36] F.-H. Tseng、J.-H. Hsueh、C.-W. Tseng、Y.-T. Yang、H.-C. Chao 和 L.-D. Chou，“利用大数据进行实时高速公路交通拥堵预测”，IEEE Access，第6卷，第57 311–57 323页，2018年。[37] X. Wang、K. An、L. Tang 和 X. Chen，“基于支持向量机和K近邻的高速公路驶出流量短期预测”，《国际交通科学与技术杂志》，第4卷，352，2015年3，第337页。–。[38] Y. Liu 和 H. Wu，“基于随机森林的道路交通拥堵预测”，载于2017年第十届计算智能与设计国际研讨会（ISCID），第2卷。IEEE，2017年，第361–364页。[39] Z. Chen、Y. Jiang 和 D. Sun，“基于时空相关性的城市道路网络交通拥堵状态辨识与预测”，IEEE Access，第8卷，第3330–3342页，2019年。[40] X. Ma、Z. Dai、Z. He、J. Ma、Y. Wang 和 Y. Wang，“将交通学习为图像：用于大规模交通网络速度预测的深度卷积神经网络”，《传感器》，第17卷，818，2017年4，第1期。页。[41] X. Ma、H. Yu、Y. Wang 和 Y. Wang，“利用深度学习理论预测大规模交通网络拥堵演化”，PloS one，第10卷，第3期，文章编号e0119044，2015年。[42] Y. Xing、X. Ban、X. Liu 和 Q. Shen，“基于对称极限学习机集群快速学习方法的大规模交通拥堵预测”，Symmetry，第11卷，第6期，文章编号730，2019年。[43] L. Lin、J. C. Handley 和 A. W. Sadek，“基于极限学习机和粒子群优化的短期交通流量区间预测”，技术报告，2017年。[44] C. Szegedy、W. Zaremba、I. Sutskever、J. Bruna、D. Erhan、I. J. Goodfellow 和 R. Fergus，“神经网络的奇特性质”，载于第二届国际学习表征会议，ICLR 2014，加拿大不列颠哥伦比亚省班夫，2014年4月14日至16日，《会议论文集》，Y. Bengio 和 Y. LeCun 编，2014年。 [在线]。可访问：http://arxiv.org/abs/
1312 6199 。S. Latif、R. Rana 和 J. Qadir，“对[45]抗性机器学习与语音情感识别：利用生成对抗网络提升鲁棒性”，arXiv 预印本 arXiv:1811.11402，2018年。[46] M. Usama、A. Qayyum、J. Qadir 和 A. Al-Fuqaha，“针对网络流量分类的黑盒对抗性机器学习攻击”，载于2019年第十五届国际无线通信与移动计算会议（IWCMC）。IEEE，2019年，第84–89页。[47] M. Usama、J. Qadir 和 A. Al-Fuqaha，“针对认知自组织网络的对抗攻击：挑战与未来方向”，载于2018年IEEE第43届局部计算机网络会议研讨会（LCN Workshops）。IEEE，2018年，第90–97页。[48] M. Usama、M. Asim、S. Latif、J. Qadir 等，“用于对网络入侵检测系统发起和阻止对抗攻击的生成对抗网络”，载于2019年第十五届国际无线通信与移动计算会议（IWCMC）。IEEE，2019年，第78–83页。[49] J. X. Morris、E. Lifland、J. Y. Yoo、J. Grigsby、D. Jin 和 Y. Qi，“TextAttack：用于NLP中对抗攻击、数据增强和对抗训练的框架”，载于《2020年自然语言处理经验方法会议：系统演示》，EMNLP 2020——演示，在线，2020年11月16日至20日，Q. Liu 和 D. Schlangen 编。计算语言学协会，2020年，第119–126页。 [在线]。可访问：https://doi.org/
10 18653 1 2020 16 。/v /。P.-Y. Chen、[50]H. Zhang、Y. Sharma、J. Yi 和 C.-J. Hsieh，“ZOO：无需训练替代模型的基于零阶优化的深度神经网络黑盒攻击”，载于第十届ACM人工智能与安全研讨会论文集，2017年，第15–26页。[51] W. Brendel、J. Rauber 和 M. Bethge，“基于决策的对抗攻击：针对黑盒机器学习模型的可靠攻击”，载于第六届国际学习表征会议，ICLR 2018，加拿大不列颠哥伦比亚省温哥华，2018年4月30日至5月3日，《会议论文集》。OpenReview.net，2018年。 [在线]。可访问：https://openreview.net/forum?id=SyZI0GWCZ [52] I. J. Goodfellow、J. Shlens 和 C. Szegedy，“解释和利用对抗样本”，载于第三届国际学习表征会议，ICLR 2015，美国加利福尼亚州圣迭戈，2015年5月7日至9日，《会议论文集》，Y. Bengio 和 Y. LeCun 编，2015年。[在线]。可访问：http://arxiv.org/abs/1412.6572 [53]A. Madry、A. Makelov、L. Schmidt、D. Tsipras 和 A. Vladu，“迈向能够抵御对抗攻击的深度学习模型”，载于第六届国际学习表征会议，*ICLR 2018*，温哥华，不列颠哥伦比亚省，加拿大，*2018*年*4*月*30*日至*5*月*3*日，会议论文集。OpenReview.net，2018年。 [在线]。参见：https://openreview.net/forum?id=rJzIBfZAb [54] N. Carlini 和 D. Wagner，“迈向神经网络鲁棒性的评估”，载于*2017*年*IEEE*安全与隐私研讨会（*SP*）。IEEE，2017年，第39–57页。[55] F. Khalid、H. Ali、H. Tariq、M. A. Hanif、S. Rehman、R. Ahmed 和 M. Shafique，“Qusecnets：基于量化的防御机制，用于保护深度神经网络免受对抗攻击”，载于*2019*年*IEEE*第*25*届在线测试与鲁棒系统设计国际研讨会（*IOLTS*）。IEEE，2019年，第182–187页。[56] H. Ali、F. Khalid、H. A. Tariq、M. A. Hanif、R. Ahmed 和 S. Rehman，“Sscnets：利用安全选择性卷积滤波器提升深度神经网络的鲁棒性”，*IEEE Design & Test*，第37卷，第2期，第58–65页，2019年。[57] N. Papernot、P. McDaniel、X. Wu、S. Jha 和 A. Swami，“蒸馏作为抵御针对深度神经网络的对抗扰动的一种防御手段”，载于*2016*年*IEEE*安全与隐私研讨会（*SP*）。IEEE，2016年，第582–597页。[58] G. S. Dhillon、K. Azizzadenesheli、Z. C. Lipton、J. Bernstein、J. Kossaifi、A. Khanna 和 A. Anandkumar，“用于鲁棒对抗防御的随机激活剪枝”，*arXiv*预印本 *arXiv:1803.01442*，2018年。[59] Y. Huang、H. Zhang、Y. Shi、J. Z. Kolter 和 A. Anandkumar，“利用高效的局部Lipschitz界训练可认证鲁棒神经网络”，神经信息处理系统进展，第34卷，2021年。[60] S. Lee、W. Lee、J. Park 和 J. Lee，“深入理解如何训练能够抵御对抗样本的可认证鲁棒模型”，神经信息处理系统进展，第34卷，2021年。[61] B. Wu、J. Chen、D. Cai、X. He 和 Q. Gu，“更宽的神经网络真的有助于提升对抗鲁棒性吗？”，神经信息处理系统进展，第34卷，第7054–7067页，2021年。[62] N. Kumari、M. Singh、A. Sinha、H. Machiraju、B. Krishnamurthy 和 V. N. Balasubramanian，“利用对抗训练模型中潜在层的脆弱性”，载于第二十八届国际人工智能联合会议，*IJCAI 2019*，澳门，中国，*2019*年*8*月*10*日至*16*日，S. Kraus 编。ijcai.org，2019年，第2779–2785页。 [在线]。参见：https://doi.org/10.24963/ijcai.2019/385