IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.3417509

Forecasting Lifespan of Crowded Events with Acoustic Synthesis-Inspired Segmental Long Short-Term Memory

Soto Anno[1], Kota Tsubouchi[2], (Member, IEEE), and Masamichi Shimosaka[1], (Member, IEEE)
[1]Tokyo Institute of Technology, Tokyo, Japan (e-mail: anno@miubiq.cs.titech.ac.jp)
[2]LY Corporation, Tokyo, Japan (e-mail: ktsubouc@lycorp.co.jp)
[3]Tokyo Institute of Technology, Tokyo, Japan (e-mail: simosaka@miubiq.cs.titech.ac.jp)

Corresponding author: Soto Anno (e-mail: anno@miubiq.cs.titech.ac.jp).

This work was partly supported by JSPS KAKENHI Grant Number 22J22725.

ABSTRACT Forecasting crowd congestion is crucial for ensuring comfortable mobility and public safety. Existing methods forecast crowding by capturing the increase in planned visits, which facilitates the methods in estimating the start of crowding. However, forecasting the change in the degree of crowding until the end is challenging owing to the lack of visitors' return plans and the deviation of visitor movements from preannounced event schedules. To address this issue, this study developed a novel framework for forecasting the start of crowding and its change over time (termed **the lifespan of crowded events (LCE)**). Based on the concept that event purposes influence the crowding patterns, our framework models these patterns according to the event purposes. Inspired by the acoustic synthesis that can successfully model the change in the sound volume for each instrument, we extended a canonical long short-term memory (LSTM) model with the concept of ADSR envelope, wherein the sound (crowd) volume changes can be represented within simple state transitions. The proposed ***versatile acoustic tri-state envelope for segmental LSTM***, namely ***VATES***, is evaluated on two datasets: synthetic and real-world mobility datasets. The results demonstrate that VATES can forecast crowding patterns with a 24.3% performance improvement, and precisely predict the start and end times of crowding, thereby improving by 6.6% and 26.1% respectively. We believe that our method enhances urban safety and mobility in crowded events, contributing to smarter city management.

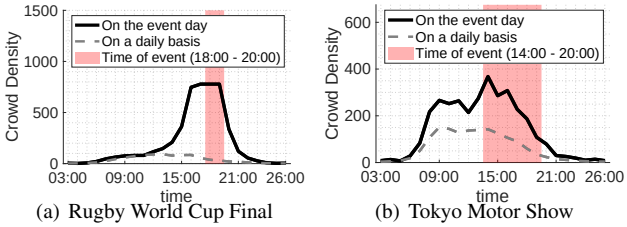
INDEX TERMS Crowd Forecasting, Urban Computing, ADSR Envelope, Acoustic Synthesis, Time Series Forecasting.

I. INTRODUCTION

Public events such as sports games, exhibitions, and festivals often result in large crowds with the potential for mobility disruptions and terrible accidents. For example, the Halloween event in Itaewon in 2022 resulted in a crowd crush that led to 159 deaths and 196 injuries[1]. Understanding the conditions for such crowded events is crucial for mitigating the risk of public threats. In particular, forecasting the **lifespan of crowded events (LCE)**, which describes the start of crowding and the change in the degree of crowding at a certain venue over time, is of great importance for several different parties: (1) such forecasting is important for event organizers and public security officials, as it enables them to allocate guards at forecasted peak congestion times, and (2) event attendees can know the forecasted end of crowding and determine the optimal time to comfortably leave the event venue.

With the extensive use of smartphones with Global Positioning System (GPS) sensors, numerous methods and applications have been developed for forecasting crowded events. These methods include simulating spatiotemporal human mobility patterns [1]–[8] and predicting the duration of crowded events [9], [10]. However, all such methods rely on the signs of congestion that appear in current mobility patterns. Therefore, the prediction time is limited to several hours ahead, which is insufficient for an initial response to crowding. Further, forecasting crowding one week ahead of the event has also been attempted. Previous research has proposed visitor schedule-driven methods that forecast the number of event visitors using their visit plans, such as transit search logs [11], [12].

[1]https://en.wikipedia.org/wiki/Seoul_Halloween_crowd_crush



(a) Rugby World Cup Final (b) Tokyo Motor Show

FIGURE 1. Transition of crowd density under various events, with the historical average (denoted as "on a daily basis") and the event hours.

However, to date, there are no methods that can forecast the LCE for two reasons. First, although visitors often plan their visits to arrive on time for an event, they seldom plan their return trips in advance. Therefore, existing methods based on the transit search logs [11], [12] can only forecast the start of crowding and not the duration and end of the crowded events. Second, event visitors do not always stay at and leave from the event venues as the event schedules announced in advance. For example, the crowd density at the Rugby World Cup started decreasing after the game finished (as shown in Fig. 1(a)), whereas that at the Tokyo Motor Show decreased during the event and immediately returned to normal after the event finished (as shown in Fig. 1(b)). Thus, forecasting the LCE is a challenging problem even when the user visit plans or event schedules are considered.

To address this issue, this study focused on the types of events, that is, event purposes, as a leading indicator of the LCE. Event visitors tend to stay at the stadiums during the sports games (Fig. 1(a)), whereas they tend to leave from the exhibition event site after finishing watching the displays they are interested in (Fig. 1(b)). Thus, the movement of people around the venue is determined by the event purposes. Consequently, events sharing the same purpose exhibit similar crowd density waveforms, whereas those with different purposes exhibit varied patterns.

Thus, our key assumption was that the waveform pattern of crowd density can be categorized by event purposes, and predicting event-purposes-based waveform patterns, such as population increase, preservation, and decrease, can facilitate more accurate LCE forecasting than predicting the crowd density for each time using a conventional time series modeling. The idea was inspired by the field of *acoustic synthesis*, where for each instrument, the shape of the volume waveform, namely *ADSR envelope* (as shown by the blue line in Fig. 2(a)), is first determined, and then a sound for a single note (as shown in the red fill in Fig. 2(a)) is synthesized. Thus, the acoustic synthesis can be viewed as generating time series of sound volume based on an approximate waveform pattern for a complete sound to provide the sound the intrinsic characteristics of instruments. Based on this concept, our hypothesis was that the LCE prediction can be rendered more accurate by learning event-purpose-based waveforms and encouraging the predicted time series of crowd density to follow the learned waveforms.

Inspired by the acoustic synthesis, this study proposed a **V**ersatile **A**coustic **T**ri-state **E**nvelope for **S**egmental LSTM, or **VATES**, to forecast the start of crowding and estimate the change in the crowd density during crowded events. We considered the event purposes as instruments, the waveform pattern of crowd density as an ADSR envelope (Fig. 2(b)), and the time series of crowd densities as a sound from onset to offset, as shown in Fig. 2(c). Based on this analogy, the VATES extended a canonical long short-term memory (LSTM) [14]-based model by leveraging the ADSR envelope concept to accurately capture waveforms tailored to specific event purposes. Specifically, we introduced an ADSR-envelope-inspired **state segmentation** strategy that extracted the state transitions and parameters of ADSR envelopes from crowd density waveforms based on event purposes. Furthermore, we introduced two auxiliary learning tasks inspired by synthesizers, in addition to the primary task of learning crowd density: (1) **envelope depiction**, to learn parameters defining the ADSR shape, and **(2) state sequence labeling**, to learn the ADSR state transitions. These auxiliary tasks involved learning the key semantics of LCE, such as the increase or decrease of crowd density, start and end times of crowding, and duration. Thus, the auxiliary tasks contribute to the improvement of LCE predictions.

The contributions of this work are as follows:

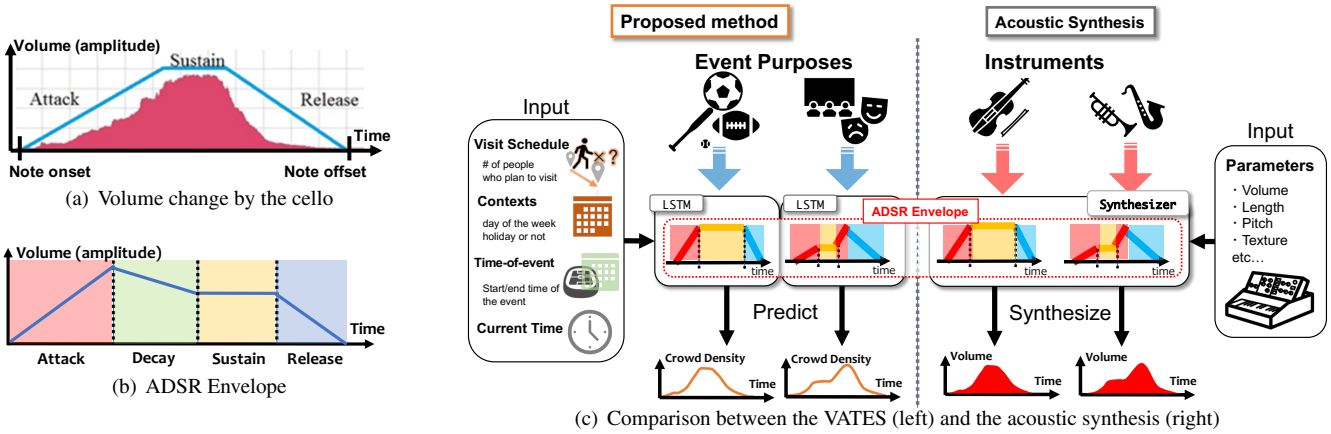
- We forecasted the LCE, that is, the start of crowding and the change in the degree of crowding over time. To the best of our knowledge, this study is the first to tackle the LCE forecast problem.
- We introduced a versatile acoustic tri-state envelope for the segmental LSTM (VATES) model, which was inspired by the successful concept of acoustic synthesis, that is, the ADSR envelope.
- The experimental results demonstrated that VATES performed better than the state-of-the-art approaches in forecasting the start of crowding and estimating time series changes in crowd densities.

The remainder of this paper is organized as follows. Section II presents a review of the related works. Section Section III presents the problem definitions and baseline approaches. Section IV describes the proposed method. Section V shows the experimental details of the performance evaluation and Section VI discusses the limitations and implications of this study. Finally, Section VII concludes the paper.

II. RELATED WORK

First, we reveal the difference between the existing **crowd density/flow prediction** and that of our study. Thereafter, we review prior studies on **urban event prediction** and **time-series forecasting**, which are related to our research. Finally, we briefly introduce literature on **acoustic synthesis**.

With the widespread use of global positioning system (GPS)-equipped devices, **crowd density and flow prediction** have been studied intensively in recent years [15], [16]. Typical studies involve simulation-based methods that assume that the spatiotemporal autocorrelation in the densities/flows of



(a) Volume change by the cello

(b) ADSR Envelope

(c) Comparison between the VATES (left) and the acoustic synthesis (right)

FIGURE 2. **(a) Sound volume changes of a single note by the cello. (b) ADSR envelope, which comprises four states: Attack (A), Decay (D), Sustain (S), and Release (R). (c) Comparison between the proposed method VATES (left) and acoustic synthesis (right). Note: The volume change patterns (a) are reprinted and adapted from [13].**

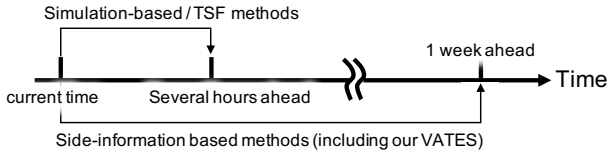


FIGURE 3. **Predictable time of the existing methods and VATES.**

people, that is, the crowd flows at a certain region and time step are correlated with the flows at the neighbor regions and next time step [1], [16], [17]. Thus, existing approaches employ current crowd flows as covariates to predict crowd flows several hours in advance [1]–[8], as shown in Fig. 3. Although these methods have exhibited promising forecasting performance, the autocorrelation assumption is easily violated when the time steps between the input and output are large. For example, signs of crowded events one week from the event do not appear in the current flow patterns. Moreover, existing methods exhibit reduced prediction performance when the time steps between the input and output are increased [4], [8].

In contrast, side-information-based methods have been proposed, wherein the crowd density/flows are predicted based on related information such as time, weather, calendar information, and visitor schedules [11], [12], [18]–[21]. Because distant future crowd patterns are predictable using related side information, this study adopted a side-information-based strategy. Typical studies have employed time, weather, and calendar information [18]–[20], where the model is specialized for predicting daily crowd densities, such as patterns in commute time. Alternatively, several studies leveraged visitor schedules that reflected when they arrived at an event venue [11], [12], [21]. However, these methods can only forecast the start of crowding as discussed in Section I. Thus, none of the existing methods can forecast the LCE one week before an event.

Urban event prediction involves the prediction of future events, such as crowd gatherings and traffic accidents in cities. Despite the difficulty in predicting such events because of their rarity, the prediction of urban events has been studied for many years [22]. Previous studies mainly focused on predicting when an event occurs by directly regressing time-to-event [23] or by fitting event time distributions with point processes [24]–[27]. However, these only predicted the occurrence of urban events, and not their duration. In contrast, several studies addressed the prediction of the duration of urban events, for example, traffic incidents [28]. Vahedian et al. [9], [10] focused on the duration of crowded events; that is, they defined the ends of crowded events as urban dispersal events and applied survival analysis to predict the length of time to the end of crowding. However, this analysis employed crowd inflows to event venues as indicators of crowding. Thus, the predictable time of this analysis was limited to the near future, that is, 5 hours ahead, in the original problem setting. In Section V, we compared our model with survival analysis-based approaches.

Time series forecasting (TSF). Herein, the typical goal is to forecast the data in the next time step based on the data in the current time step and has been extensively researched [29]. The TSF technique has been applied in many fields, such as climate modeling [30], biological sciences [31], and finance [32]. With the emergence of deep learning technology in recent years, recurrent neural networks (RNN), such as LSTM [14] and Gated Recurrent Unit (GRU) [33], are frequently used to model time-series data. Furthermore, convolutional neural networks (CNN) [34], which were originally introduced in the computer vision field, have gained attention as a TSF technique [35], [36], particularly for modeling spatiotemporal abnormal events [37]. Although promising methodologies have been proposed in the TSF literature, the underlying assumption in the existing TSF framework involves the capture of the temporal autocorrelation between the input and output data, which is akin

to the simulation-based methods for crowd flow prediction. Thus, crowding one week in advance cannot be forecasted, as shown in Fig. 3. Although we adopted LSTM to capture the time series of the LCE, we combined the side-information-based strategy with the LSTM-based model to realize one-week-ahead forecasting.

Acoustic synthesis is a fundamental technology in modern popular music [38]–[40]. IN general, synthesizers use an envelope that governs the time-variant volume change of a note from onset to offset to imitate the instruments. One of the most prevalent envelopes, ADSR (as shown in Fig. 2(b)), comprises four states that define the volume changes: attack (a rising phase of sounds), decay (first attenuation phase of sounds), sustain (preserving phase of sounds), and release (final attenuation phase of sounds) [41]. To the best of our knowledge, this is the first study to use the ADSR approach for predicting urban events. Specifically, we applied the ADSR approach to synthesize the forecasted crowd density transition.

III. PRELIMINARIES

A. PROBLEM SETTING

Let l be an event venue and d represent an index of dates. We divided a day into T time segments denoted by t (i.e., $t := 1, ..., T$). Further, $\tau := dT + t$ denotes the time steps indexed in the sequential order of the time series within a dataset. The *crowd density* observed at venue l at time step τ is denoted by $y_{l,\tau}$.

We modeled $y_{l,\tau}$ by leveraging the side information such as the time segments of the day, contextual information (i.e., day of the week and holidays), pre-scheduled visitor counts for venue l, event hours, and event purpose (e.g., sports and exhibitions). Let $\boldsymbol{t} \in \mathbb{R}^T$ be a time feature of time segment t, and $\boldsymbol{c}_d \in \mathbb{R}^C$ be a context feature encoding the day of the week or holiday. To estimate the visitor count of events accurately, we incorporated the number of scheduled visits by attendees. For this purpose, we used transit search logs based on previous studies [11], [12], [21]. The transit search log is a tuple (d, t, d', l) of scheduled date d, time t, search date d', and destination l. The number of scheduled visits for l on date d as of d' is accounted for by the number of logs that schedule visits to destination l on date d and are searched on date d', denoted by $x_{l,d,t|d'}$. Subsequently, a scheduled visit feature is defined as $\boldsymbol{x}_{l,d} = \{x_{l,d,j|d-i} | i = p_d, ..., p_d + p_w, j = 1, ..., T\} \in \mathbb{R}^{p_w T}$, where p_d denotes the earliest day preceding the scheduled date d and p_w represents the span of the days under consideration. To capture the duration of crowding, we leveraged event-hour data. We formulated a time-of-event feature $\boldsymbol{e}_{l,d} \in \mathbb{R}^E$ that represented the start and end times of an event on day d. We further considered the event purpose, such as sports games and exhibitions, denoted by $\boldsymbol{u}_{l,d} \in \mathbb{R}^U$. A comprehensive formulation of these features is presented in Section V-B1.

We forecasted the LCE at l by learning and forecasting the time series of $y_{l,\tau}$ *one week before an event*. Specifically, considering the side information such as scheduled visit feature $\boldsymbol{x}_{l,d}$, contextual feature $\boldsymbol{c}_d$, time feature $\boldsymbol{t}$, time-of-event feature $\boldsymbol{e}_{l,d}$, and event purpose $\boldsymbol{u}_{l,d}$, we established a predictive model of $y_{l,\tau}$ for each venue l, as of $d - 7$.

B. BASELINE: CANONICAL LSTM WITH EVENT PURPOSES

The forecasting of the $y_{l,\tau}$ can be regarded as the TSF problem. However, as discussed in Section II, the existing TSF model assumes the temporal autocorrelation, which limits the predictable time to several hours, as shown in [4], [8]. Therefore, we incorporated the side information (such as event purposes) based strategy with the canonical Long Short-Term Memory (LSTM) framework [14] as the baseline model.

As we can assume that the number of GPS logs within a certain region and time interval follows a Poisson distribution [18]; Thus, the likelihood of $y_{l,\tau}$ can be expressed as $\mathcal{P}(y_{l,\tau}) = \mathrm{Pois}(y_{l,\tau}|\lambda_{l,\tau}) = {\lambda_{l,\tau}}^{y_{l,\tau}} \exp(-\lambda_{l,\tau})/y_{l,\tau}!$, where $\lambda_{l,\tau} > 0$ denotes the mean parameter of the Poisson distribution.

Further, $\lambda_{l,\tau}$ was regressed using the input features introduced in Section III-A. Specifically, we generated an embedding vector $\boldsymbol{h}_\tau^{(0)}$ from the features as inputs to the canonical LSTM model. To handle the coupling effect between time and other factors, we adopted a bilinear representation as follows:

$$\boldsymbol{o}_d^{(0)} := \mathrm{Concat}(\boldsymbol{c}_d, \boldsymbol{x}_{l,d}, \boldsymbol{e}_{l,d}, \boldsymbol{u}_{l,d}) \in \mathbb{R}^{C+p_w T+E+U}, \quad (1)$$

$$h_{\tau,i}^{(0)} := \mathrm{ReLU}({\boldsymbol{o}_d^{(0)}}^\top \boldsymbol{W}_i^{(0)} \boldsymbol{t}),\ i = 1, ..., H, \quad (2)$$

$$\boldsymbol{h}_\tau^{(0)} := [h_{\tau,1}, ..., h_{\tau,H}]^\top \in \mathbb{R}^H, \quad (3)$$

where $\mathrm{Concat}(\cdot)$ is a column-oriented concatenation operation of vectors and $\boldsymbol{W}_i^{(0)} \in \mathbb{R}^{(C+p_w T+E+U)\times T}$ is a matrix of parameters. The LSTM-based model parameterizes the time series of $\lambda_{l,\tau}$ as follows:

$$\boldsymbol{h}_\tau, ..., \boldsymbol{h}_{\tau+p} := \mathrm{LSTM}(\boldsymbol{h}_\tau^{(0)}; \boldsymbol{\Theta}_{\mathrm{LSTM}}), \quad (4)$$

$$\ln \lambda_{l,\tau+j} = \boldsymbol{w}_j^\top \boldsymbol{h}_{\tau+j},\ j = 0, ..., p, \quad (5)$$

where $\boldsymbol{\Theta}_{\mathrm{LSTM}}$ and $\boldsymbol{w}_j \in \mathbb{R}^H$ are the learning parameters, and $\boldsymbol{h}_{\tau+j} \in \mathbb{R}^H$ denotes the hidden feature as interpreted by the LSTM model. The loss function is the negative log-likelihood (NLL) of the Poisson distribution, defined as $\mathcal{L}^{(l)}(\boldsymbol{\Theta}_{\mathrm{LSTM}}) = -\sum_\tau \sum_{j=0}^p \ln \mathrm{Pois}(y_{l,\tau+j}|\lambda_{l,\tau+j})$.

This model takes into account the event purposes that are essential for distinguishing waveform differences associated with various events. However, merely considering these purposes are not sufficient to accurately model the waveforms because these purposes do not include detailed waveform semantics such as peak crowd density and crowding duration.Therefore, in the next section, we will introduce a novel approach to waveform modeling that utilizes the successful techniques of acoustic synthesis.

IV. PROPOSED METHOD: VATES

A. BASIC IDEA OF VATES:

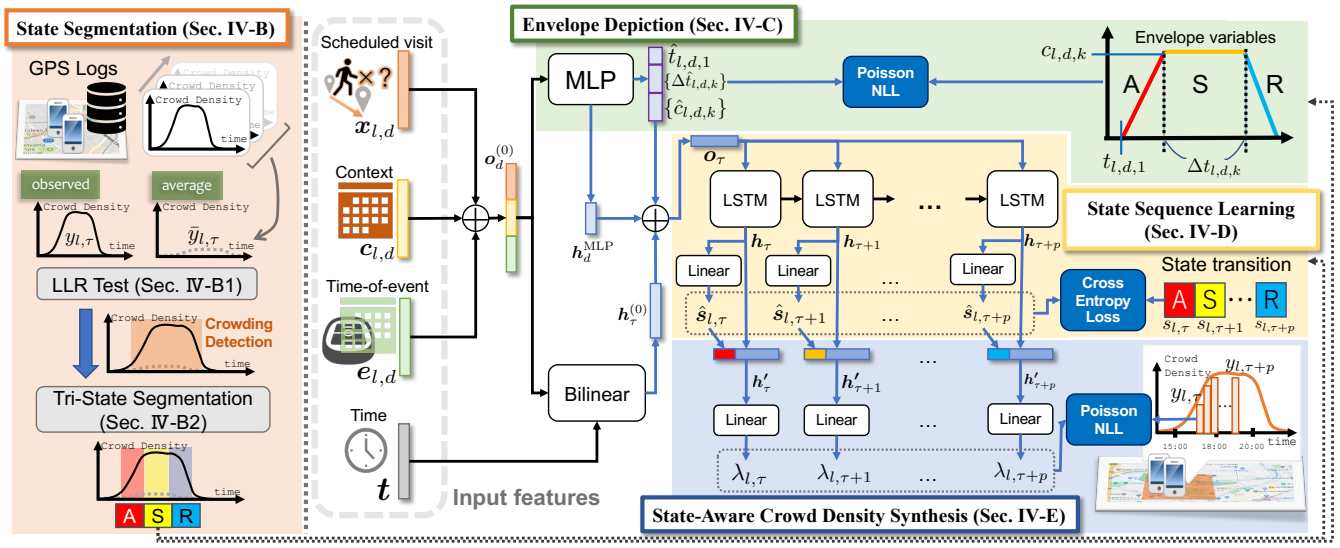


FIGURE 4. **Overview of VATES. The section number which describes each component is indicated in the figure.**

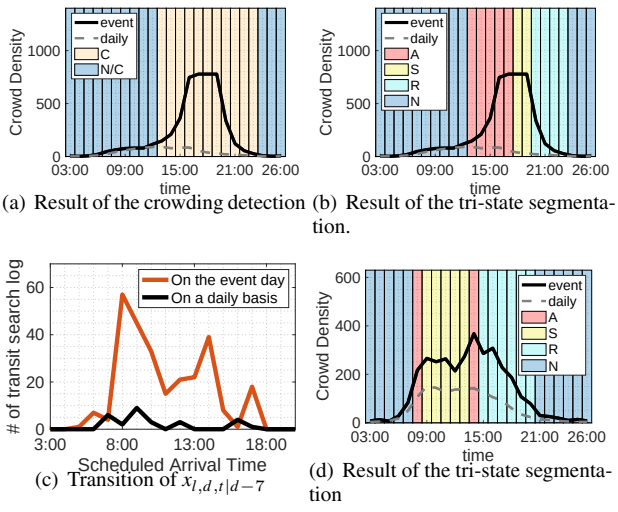
INSPIRATION FROM ACOUSTIC SYNTHESIS

The basic idea of the versatile acoustic tri-state envelope for segmental LSTM (VATES) is to first predict the shape of the crowd density transition by the event purpose, instead of directly regressing the crowd density for each time step. The shape reflects the inherent characteristics of crowded events, that is, the timing of the start and end of congestion, and the presence of periods in which the number of visitors increase or decrease. This approach is analogous to acoustic synthesis, in which the envelope shape is first determined such that it matches the desired instrument.

Thus, we incorporated the concept of ADSR envelope into the baseline canonical LSTM model introduced in Section III-B. From the ADSR envelope, we adopt the following three states: attack (increase in crowd density), sustain (preservation of density), and release (decrease in density). We also introduced a noncrowded state to represent a state other than a crowded event. As illustrated in Fig. 4, the proposed VATES extends the baseline canonical LSTM model with the following components:

1) **State Segmentation** — This component categorizes the events by event purposes and segments $y_{l,\tau}$ in the following four states: attack, sustain, release, and non-crowded (discussed in Section IV-B).
2) **Envelope Depiction** — This component is to learn and predict the shape of crowd densities that are used to synthesize the prediction of crowd density $\hat{y}_{l,\tau}$ (discussed in Section IV-C).
3) **State Sequence Labeling** — This component learns and predicts the states of each time step ($\tau \sim \tau + p$) based on the predicted envelopes, thereby assisting in forecasting crowd densities associated with state transitions (Section IV-D).
4) **State-Aware Crowd Density Synthesis** — Drawing on the predicted envelope and state sequences, this component synthesizes the prediction of crowd density (Section IV-E).

B. STATE SEGMENTATION



(a) Result of the crowding detection (b) Result of the tri-state segmentation.

(c) Transition of $x_{l,d,t|d-7}$ (d) Result of the tri-state segmentation

FIGURE 5. **(a) Crowding detection result and (b) Tri-state segmentation result for Rugby World Cup Final. C denotes "overcrowded," and N/C denotes "daily". (c) Transition of the number of transit search logs $x_{l,d,t|d-7}$, where l is Tokyo Big Sight where the Tokyo Motor Show was held, and d is Oct. 25, 2019. (d) Tri-state segmentation result for Tokyo Motor Show.**

This component is responsible for segmenting $y_{l,\tau}$ into four distinct states: attack, sustain, release, and non-crowding. We first divided the data into crowded (including attack, sustain, and release) and noncrowded, as discussed in Section IV-B1.

We then segmented the data into attack, sustain, and release by considering the event purposes discussed in Section IV-B2.

1) Crowding Detection by LLR test

We first divided the data into crowded and non-crowded states. However, because the magnitude of congestion fluctuates between event venues, dividing raw data into crowded and noncrowded is nontrivial. To address this issue, we propose a likelihood ratio (LLR) test-based crowding detection, a statistical method for detecting deviations from historical norms.

Let $\bar{y}_{l,\tau}$ be the daily density calculated using the historical average for the same day of the week and time. Next, we investigate whether $y_{l,\tau}$ is significantly greater than $\bar{y}_{l,\tau}$. Assuming that $y_{l,\tau}$ follows the Poisson distribution, we test the following hypotheses: $H_0 : y_{l,\tau} \sim \mathrm{Pois}(\cdot|\bar{y}_{l,\tau})$, $H_1 : y_{l,\tau} \sim \mathrm{Pois}(\cdot|\bar{z}_{l,\tau})$ where $\bar{z}_{l,\tau}$ is the mean parameter that satisfies $\bar{z}_{l,\tau} > \bar{y}_{l,\tau}$. To execute this test, we employ the Expectation-based likelihood ratio test proposed by Neill [42], where the test statistic is the likelihood ratio $\mathrm{LLR}_l(\tau)$, which was formulated as follows:

$$\mathrm{LLR}_l(\tau) = \begin{cases} y_{l,\tau} \log \frac{y_{l,\tau}}{\bar{y}_{l,\tau}} + (\bar{y}_{l,\tau} - y_{l,\tau}) & \text{if } y_{l,\tau} \geq \bar{y}_{l,\tau}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

$y_{l,\tau}$ can be classified as crowded if $\mathrm{LLR}_l(\tau) > 0$ and $\mathrm{LLR}_l(\tau)$ is significant at α-level. Although determining the statistical significance of a likelihood ratio generally requires computationally expensive Monte Carlo methods, $\mathrm{LLR}_l(\tau)$ is significant at α-level if and only if $1 - \Pr(X < y_{l,\tau}) \geq \alpha$ as demonstrated by Zhou et al. [43]. Fig. 5(a) provides the result of crowded (C) and non-crowded (N/C) density extractions.

2) Tri-State Segmentation by Event Purposes.

The extracted densities in the crowded state were further segmented into a three tri-state: attack, sustain, and release. Segmentation was performed by the event-purpose-by way to obtain a shape that reflects the heterogeneity of the event purposes. Thus, we categorized crowded events into several types and introduced segmentation methodologies customized for these purposes. In this study, we introduced the following two types of event purpose:

- **sports-type** — These events have an attack, sustain, and release sequence, each phase appearing independently. Typically hosted in stadiums or arenas, this category comprises events, such as sporting matches (illustrated in Fig. 1(a)).
- **exhibition-type** — These events have a double-attack sequence, in the order of attack, sustain, attack, release. Generally organized in *exhibition halls*, this category includes large-scale expositions, such as Tokyo Motor Show shown in Fig. 1(b).

Customizing the segmentation method for each event purpose is similar to defining the ADSR envelope for each musical instrument. VATES can be extended to other event purposes that are not addressed in this study (e.g., firework displays and festivals) if the combination of the three states is determined.

In sports-type events, such as sports or concert events, we observed the following: (1) entrance and exit times are explicitly set by event timings. (2) Spectators stay at the venue throughout the event. Thus, we segmented the crowding densities as follows:

1) **attack**: From the start of crowding to event start.
2) **sustain**: During the event.
3) **release**: From the end of event to the end of crowding.

Fig. 5(b) illustrates the segmented results on a sports-type event.

Exhibition-type events, such as large-scale expositions staged in exhibition halls, exhibit congestion patterns distinct from sports-type events. The observed characteristics were as follows:

- Queue-induced crowding occurs before doors open, due to exhibit enthusiasts arriving early for priority access. An increase in crowd density is observed before the event starts (6:00-9:00 in Fig. 1(b)). Moreover, these enthusiasts often plan their transits in advance, resulting in an increase in transit searches targeting this period, as shown in Fig. 5(c).
- The number of visitors increases again after the event starts because non-enthusiast participants arrive (14:00-15:00 in Fig. 1(b)).
- Majority of visitors refrain from staying until the event termination, leaving before the event ends (15:00-20:00 in Fig. 1(b)).

Based on these observations, exhibition-type events have a double attack sequence in the order of attack, sustain, attack, and release. Segmentation of the extracted crowd densities in exhibition halls was performed by leveraging scheduled visits and event timings as follows:

1) **first attack:** From the crowding start to the peak of scheduled visits.
2) **sustain:** Between the first and second attacks.
3) **second attack:** From event start to the peak of crowd density.
4) **release:** From crowd density peak to the end of crowding.

Fig. 5(d) shows the segmented results on an exhibition-type event.

C. ENVELOPE DEPICTION

This component learns and predicts the shape of the envelope used to predict the crowd density $\hat{y}_{l,\tau}$. To learn the shape, we modeled the following variables: (1) the start time of the first attack state, (2) the time length of each state, and (3) the crowd densities at the end of each state. We built a multilayer perceptron (MLP) model to model these variables. For the input of the MLP, we used the context, scheduled visit, and time-of-event features.

Let $t_{l,d,k} \in \{1, ..., T\}$ be the start time of the k-th state and $\Delta t_{l,d,k} \in \{1, ..., T\}$ be the time length of the k-th state on the event date d. Furthermore, $c_{l,d,k}$ is the crowd density at the end of the kth state. If the kth state is released, we set $c_{l,d,k} =$

0; otherwise, we set $c_{l,d,k} = y_{l,\tau}$, where $\tau = t_{l,d,k+1} - 1$. Because these values are all nonnegative discrete values, we assumed that they follow individual Poisson distributions and formulated the MLP as follows:

$$\begin{aligned} \boldsymbol{h}_d^{\mathrm{MLP}} &:= \mathrm{MLP}(\boldsymbol{c}_d, \boldsymbol{e}_{l,d}, \boldsymbol{x}_{l,d}, \boldsymbol{u}_{l,d}; \boldsymbol{\Theta}_{\mathrm{MLP}}), \quad (7) \\ \ln \pi(\beta)_d &:= \boldsymbol{w}_\beta^\top \boldsymbol{h}_d^{\mathrm{MLP}}, \\ \text{where } \beta \in \mathcal{A}_d &:= \{t_{l,d,1}\} \cup \{\Delta t_{l,d,k}\}_{k=1}^K \cup \{c_{l,d,k}\}_{k=1}^K, \quad (8) \end{aligned}$$

where $\pi(\beta)_d$ is the mean parameter of the Poisson distribution whose β follows, $\boldsymbol{h}_d^{\mathrm{MLP}} \in \mathbb{R}^H$, and $\boldsymbol{w}_\cdot \in \mathbb{R}^H$, and K is the number of states.

D. STATE SEQUENCE LABELING

This component learns and predicts the states for the target time steps ($\tau \sim \tau + p$) from the envelope variables predicted by the MLP described in Section IV-C. As changes in crowd density are associated with segmented state transitions, acquiring this knowledge can facilitate the capture of density shifts under crowding conditions. Let $\boldsymbol{s}_{l,\tau} \in \mathbb{R}^4$ be a one-hot vector for a state label $s_{l,\tau} \in \mathcal{S} := \{\mathrm{A}, \mathrm{S}, \mathrm{R}, \mathrm{N}\}$ in a time step τ, wherein A, S, R, and N are attack, sustain, release, and non-crowded, respectively. Building on Softmax-based classification, we extended the baseline LSTM model by integrating the envelope variables as inputs as follows:

$$\begin{aligned} \boldsymbol{o}_\tau &:= \mathrm{Concat}(\boldsymbol{h}_\tau^{(0)}, \boldsymbol{h}_d^{\mathrm{MLP}}, \{\pi(\beta)_d\}_{\beta \in \mathcal{A}_d}) \\ &\quad \in \mathbb{R}^{2H+2K+1}, \quad (9) \\ \boldsymbol{h}_\tau, ..., \boldsymbol{h}_{\tau+p} &:= \mathrm{LSTM}(\boldsymbol{o}_\tau;\ \boldsymbol{\Theta}_{\mathrm{LSTM}}), \quad (10) \\ \hat{\boldsymbol{s}}_{l,\tau+j} &:= \log(\mathrm{Softmax}(\boldsymbol{W}_j \boldsymbol{h}_{\tau+j})),\ j = 0, ..., p, \quad (11) \end{aligned}$$

where $\hat{\boldsymbol{s}}_{l,\tau+j} \in \mathbb{R}^4$ represents the logit for classification and $\boldsymbol{W}_j \in \mathbb{R}^{4 \times H}$ is the learning parameter.

E. STATE-AWARE CROWD DENSITY SYNTHESIS

Based on the envelope variables predicted in Section IV-C and the state sequence predicted in Section IV-D, this component synthesized crowd density $\hat{y}_{l,\tau}$. We reformulated the mean parameter of the Poisson distribution for $y_{l,\tau}$ as follows:

$$\begin{aligned} \boldsymbol{h}'_\tau &:= \mathrm{Concat}(\boldsymbol{h}_\tau, \hat{\boldsymbol{s}}_{l,\tau}), \quad (12) \\ \ln \lambda_{l,\tau+j} &:= \boldsymbol{w}_j^\top \boldsymbol{h}'_{\tau+j},\ j = 0, ..., p, \quad (13) \end{aligned}$$

where $\boldsymbol{w}_j^\top \in \mathbb{R}^{4+H}$ denotes the learning parameter.

Given the ground-truth crowd density $y_{l,\tau}$, envelope variables $\beta \in \mathcal{A}_d$, and state labels $s_{l,\tau}$, our task can be regarded as multi-task learning that incorporates Poisson regression and multi-class classification problems. Consequently, we derived the following loss function:

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\Theta}_{\mathrm{LSTM}}, \boldsymbol{\Theta}_{\mathrm{MLP}}, \boldsymbol{W}_{i:H}^{(0)}, \boldsymbol{W}_{0:p}, \boldsymbol{w}_{0:p}, \boldsymbol{w}_\beta) \\ &= -\sum_l \sum_d \left[\sum_t \sum_{j=0}^p (\mathcal{L}_y(y_{l,\tau+j}) + \mathcal{L}_s(\boldsymbol{s}_{l,\tau+j})) + \mathcal{L}_\beta(\alpha_d) \right], \quad (14) \end{aligned}$$

where

$$\begin{aligned} \mathcal{L}_y(y_{l,\tau+j}) &= \ln \mathrm{Pois}(y_{l,\tau+j} | \lambda_{l,\tau+j}), \quad (15) \\ \mathcal{L}_s(\boldsymbol{s}_{l,\tau+j}) &= \sum_{i=1}^{|\mathcal{S}|} \{\boldsymbol{s}_{l,\tau+j}\}_i \log\{\hat{\boldsymbol{s}}_{l,\tau+j}\}_i, \quad (16) \\ \mathcal{L}_\beta(\alpha_d) &= \sum_{\beta \in \mathcal{A}_d} \ln \mathrm{Pois}(\alpha_d | \pi(\beta)_d). \quad (17) \end{aligned}$$

$\mathcal{L}_y(y_{l,\tau+j})$ and $\mathcal{L}_\beta(\alpha_d)$ are the NLL of the Poisson distribution, and $\mathcal{L}_s(\boldsymbol{s}_{l,\tau+j})$ is the cross-entropy loss. Notably, this model can be trained end-to-end.

V. EXPERIMENTS

VATES was empirically evaluated through the following experiments.

- **Forecasting Performance Evaluation** (Section V-C): We evaluated the forecasting performance in crowd density, start, and end times of crowding in various events with both synthesized and real data.
- **Qualitative Prediction Evaluation** (Section V-D): We provided case studies with visual representations of forecasted crowding during actual events.
- **Crowding Synthesis with Manipulated Envelopes** (Section V-E): We demonstrated that the crowd density transition predicted by our synthesizer-inspired framework can be controlled when we manipulate the envelope externally.
- **Ablation Study** (Section V-F): We identified the source of enhancement in the forecasting performance of VATES.

A. DATASET

Although there are various open datasets that are often used for modeling the human mobility or crowd flows, such as Yellow Taxis in New York City [2] and BikeNYC[3], all of these datasets do not contain the people's visit schedules such as transit search logs. To the best of our knowledge, there is no open dataset that is applicable to the LCE forecast. Thus, we created synthetic data that behaved the LCE and collected real data from large-scale real-world events.

1) Synthetic Data

We experimented on synthesized sports- and exhibition-type event data from Poisson models as follows: $y_\tau := \bar{y}_\tau + \Delta y_\tau$, where $\bar{y}_\tau$ is the daily density sampled from a Poisson model $\mathrm{Pois}(\cdot | \exp(\hat{\boldsymbol{c}}_d^\top \hat{\boldsymbol{W}} \boldsymbol{t}))$ with $\hat{\boldsymbol{c}}_d \in \mathbb{R}^C$ sampled from

[2] https://opendata.cityofnewyork.us/overview/
[3] https://citibikenyc.com/

TABLE 1. Summary of four large-scale events used in an empirical evaluation.

Event Name	Rugby World Cup Final	J1 League Final Section	Comic Market	Tokyo Motor Show
Overview	Rugby Match	Soccer Match	Self-Published Comic Show	Auto Show for Cars, etc.
Date	11/2, 2019	12/7, 2019	12/28 ∼ 31, 2019	10/24 ∼ 11/4, 2019
# of visitors	70,103	63,854	750,000 in total	1,300,900 in total
Event Venue	Nissan Stadium (https://www.nissan-stadium.jp/)		Tokyo Big Sight (https://www.bigsight.jp/)	
Event Type	**sports**		**exhibition**	

a uniform distribution $\mathrm{Uni}(0,1)$ and $\hat{\boldsymbol{W}} \in \mathbb{R}^{C\times T}$ sampled from $\mathrm{Uni}(0,0.04)$, and Δy_τ is the surge in density during crowded events, sampled from another Poisson model $\mathrm{Pois}(\cdot|\mu_\tau)$ where μ_τ is changed according to the state transition. The start time of events t_{start} was randomly sampled between 8 am and 2 pm, and the time length of the event was randomly sampled between 5 and 10 h.

To simulate sports-type events, we categorized the time period during an event as Sustain, and μ_τ was set to $\mu_{\mathrm{peak}} \in [250, 500]$ throughout Sustain. The attack and release durations were two if $\mu_{\mathrm{peak}} < 300$, three if $300 \leq \mu_{\mathrm{peak}} < 400$, and five otherwise. μ_τ is defined as zero during noncrowded conditions, linearly increases from zero during attack, and linearly decreases to zero during release. Moreover, we generated a scheduled visit feature $\boldsymbol{x}_d$ with a Gaussian kernel as $\boldsymbol{x}_d = \{t\mu_{\mathrm{peak}}|t = \mathcal{N}(j|t_{\mathrm{start}}, \sigma^2), j = 1, ..., T\}$.

To simulate exhibition-type events, we denoted the start time of the second attack as t_{start}, and its duration as one if $\mu_{\mathrm{peak}} < 300$, two if $300 \leq \mu_{\mathrm{peak}} < 400$, and three otherwise, after which the time until the event ends is categorized as release. Before the second attack, we introduced the first attack and sustain, where the duration of the first attack was the same as that of the second attack, and the duration during sustain was $\{1, 2, 3\}$. We assigned μ_τ to a fixed $\mu_{\mathrm{peak}} \in [250, 500]$ at the end of the second attack, and $\mu_\tau = \mu_{\mathrm{peak}}/2$ during sustain.

For the synthetic data, we generated 180 days of training data, within which 1% (equating to 1 ∼ 2 days) were designated event dates and the remaining 99% of the training data was non-crowded. This aimed to simulate the situation where the event inducing the crowding is infrequent and abnormal, as discussed in Section I. Additionally, we generated 180 days of event data for evaluation.

2) Real Data: GPS logs, Transit Search logs, Event Calendar

To evaluate VATES based on real data, we collected GPS-based mobility logs, transit search logs and event calendar information. The GPS logs were procured using a mobile application from LY Corporation on October 1st, 2019 ∼ February 28th, 2020. Each record, collected with user consent, was entirely anonymized by replacing the user IDs with dummy identifiers and was characterized by timestamp, latitude, and longitude. We aggregated the mobility logs within each event venue, demarcated by a 500 × 500 m square area at each time segment, and tabulated their quantity as crowd density. Hence, we refrained from using any dataset with personally identifiable information for the data analysis and model construction. For the scheduled information of the user, we also used transit search logs, which were searched mainly by train passengers. These logs were assembled using the transit search engine[4], which was released by LY Corporation. Each record contained an anonymized user ID, search timestamp, scheduled timestamp, and destination station. Analogous to the mobility logs, we enumerated the volume of search records per station and time segment, thereby avoiding the use of personal information. We also employed event calendar data for time-of-event information. Each record comprises the date, event name, start time, and end time. We formulated time-of-event features based on these data.

Our model was evaluated for four large-scale, non-regular, and abnormally crowded events staged in Japan on October 1st, 2019 ∼ February 28th, 2020, as presented in Table 1. We used each event as the evaluation data, and the rest of the data that shared the same event purpose was used as the training data. The J1 League Final Section, typically a weekly event, was considered an abnormally crowded event because it attracted the largest recorded attendance.

B. EXPERIMENTAL SETTINGS

1) Model and Feature Settings

We considered one day as a 24-h period, and the number of time segments T was set to 24 (i.e., one time segment denotes a 1-h period). Following previous research [18], the start of the day was at 3:00 a.m., which had the least active population, and the end was at 3:00 a.m. the next day (i.e., 27:00 in 24-h notation). To execute the forecasting as of $d-7$, we set $p_d = 7$. We also allocated $p_w = 7$ to consider the schedule patterns specified two weeks before the event day. For the context feature $\boldsymbol{c}_d$, we use the days of the week, holidays, and weekdays or weekends. By implementing one-hot encoding, the days of the week became a seven-dimensional vector, whereas holiday-or-not and weekday-or-weekend were two-dimensional vectors. By employing a tensor product to compose these features, we obtained $\boldsymbol{c}_d \in \mathbb{R}^{28}$. We formulated the time feature as $\boldsymbol{t} = \{u|u = \mathcal{N}(j|t, \sigma^2), j = 1, ..., T\}$ by following [18]: For the time-of-event feature, we encode the event times with multi-hot encoding, yielding $\boldsymbol{e}_d^{\mathrm{ED}} = \{\mathbb{I}[t \in \{t_{\mathrm{on}}, t_{\mathrm{off}}\}]|t = 0, ..., T\} \in \mathbb{R}^T$, where t_{on} is the start end t_{off} is the end time of the event. For the event purpose feature denoted as $\boldsymbol{u}_{l,d}$, we use one-hot encoding of *sports* and *exhibition*, which yields $\boldsymbol{u}_{l,d} \in \mathbb{R}^2$. In VATES, we use a four-layer LSTM and MLP with 256-dimensional outputs in the hidden space, followed by ReLU activation. We set $p = 12$

[4] https://transit.yahoo.co.jp/

and $\alpha = 1 \times 10^{-6}$. We adopted a batch size of 64 and used the Adam optimizer with a learning rate 2×10^{-4}.

The training data were separated in a 9:1 ratio, with the major and minor segments used as the initial training data and validation, respectively. Early stopping was used in the model training, whereby the model was initially trained solely on the initial training data and the training was repeated until the loss in the validation data stopped improving over 10 epochs. The model was then trained for 10 epochs on all training data, including the initial training and validation data.

2) Evaluation Metrics

To evaluate the forecasting performance of the LCE, we adopt three error metrics: (1) mean absolute error (**MAE**) for crowd densities within each state, namely, attack (A), sustain (S), release (R), and non-crowded (N), (2) mean absolute starting time error (**MASTE**), signifying error in crowding starting time, and (3) mean absolute ending time error (**MAETE**) representing error in the ending time of crowding. To calculate the MASTE and MAETE, the start and end times of the crowding have to be predicted. To predict them, we applied the LLR test presented in Section IV-B1 to the prediction results of the crowd densities, where the minimum time of the crowding period was regarded as the start time, while the maximum time of the crowding period was regarded as the end time. LLR can fail to detect crowded density if the predictions underestimate the ground truth at α-level. When we evaluated the performance of the synthetic data, treated the failure case of LLR as NaN, filtered out the NaN result, and calculated MASTE and MAETE from the remaining test data. We further tested the accuracy (**Acc.**) for evaluating the number of success and failure cases in the LLR test for each method in the synthetic data experiment. When evaluating the performance on real data, we denoted it as NaN in the experimental results. Note that MAEs under each state were not NaN, and we evaluated them on all data without filtering.

3) Comparison Methods

Firstly, we compared VATES with the side-information-based methods for forecasting crowding including a state-of-the-art approach as follows:

- **Event-aware Historical Average (EHA)**: This calculates the average values corresponding to the same time segments on the day when an event was held within the training data.
- **Bilinear Poisson Regression (BPReg*)** [18]: A regression model formulated to use side information as input features. While the original work adopted only the contextual and time features, we expanded its usage of side information to incorporate the scheduled visit and time-of-event features.
- **CityOutlook** [12]: A state-of-the-art crowding forecasting method using side information including visitors' schedules. Following the original work, this method adopts contextual, time, and scheduled visit features.
- **Non-Segmental LSTM (NS LSTM, Section III-B)**: A canonical LSTM model, introduced as a baseline method of VATES in Section III-B. This model simply considers the event purposes but does not employ the proposed synthesizer-inspired technique.

Note that these baselines were trained individually for each event purpose, which was also adopted in VATES.

Additionally, we compared VATES with the survival analysis (SA) [10]-based method, which is the state-of-the-art approach for forecasting the duration of crowding. We compare VATES with the following SA-based methods.

- **Survival Analysis (SA)** [10]: A state-of-the-art method for forecasting the duration of crowding. This method forecasts the end time of crowding after its start. While we used the original profiles as input features except POI vectors, we replaced the counts of taxi pick-up and drop-off with the crowd density data for the recent and target profiles. We did not use POI vectors because all the target POI in this study was the event venue. We evaluated the performance in the original setting, that is, **5-h ahead of prediction**.
- **SA + {BPReg, CityOutlook, VATES}**: This is the extension of survival analysis-based prediction [10], where the input crowd density features of the method are replaced by the forecasted results of BPReg, CityOutlook, and VATES in the testing phase. We evaluated the performance of this method on **the one week ahead of forecasting**.

We did not compare VATES with the existing simulation-based crowd flow prediction methods such as [3], [4], [7], [8]. This is because we are interested in the persistence of crowding at the event venue one week ahead, which is orthogonal to the spatiotemporal crowding occurrence several hours ahead that was addressed by the simulation-based methods.

We implemented VATES and the comparison methods with the MapReduce framework implemented in Apache Spark [44] and the BigDL framework [45] for our empirical evaluation.

C. FORECASTING PERFORMANCE EVALUATION

Table 2 and 3 list the comparative performance analysis of VATES and the baselines using synthetic and real data, respectively. Note again that NaN in Table 3 signifies instances in which the LLR test failed to detect crowding owing to the underestimation of crowd densities. For the definition of the used error metrics, see Section V-B2. On average, the results show that VATES improved the prediction performance by 24.3% in MAE, 6.6% in MASTE, and 26.1% in MAETE, respectively, compared with the state-of-the-art methods.

From the results of the synthetic data in Table 2, although EHA showed the minimum errors in MAE for the N state, EHA showed the poorest performance in MAE across all ASR states, MASTE, MAETE, and Acc. CityOutlook achieved a relatively small MAE on the A state for the sports-type, indicating its capabilities to capture the start of crowding. However, it failed to show satisfactory performance on MAE for the S and R states, MASTE, and MAETE. However, VATES outperformed all baselines designed for one-week-ahead forecasting, indicating the effectiveness of the proposed envelope-based strategy.

TABLE 2. Performance comparison of VATES and Baselines in MAE for each state, MASTE, MAETE, and Accuracy (Acc.) on synthetic data. The unit of MAE, MASTE/MAETE, and Acc. are the number of GPS logs, hours, and %, respectively.

Model	sports-type							exhibition-type						
	MAE↓				MASTE↓	MAETE↓	Acc.↑	MAE↓				MASTE↓	MAETE↓	Acc.↑
	A	S	R	N				A	S	R	N			
EHA	246.2	368.2	147.7	**18.8**	23.7	23.6	1.7	406.6	526.9	273.6	**17.7**	23.7	23.6	1.7
BPReg* [18]	229.2	307.1	199.0	22.2	2.1	1.3	98.5	275.7	357.9	219.8	32.9	3.4	2.5	98.5
CityOutlook [12]	124.1	248.1	115.4	58.9	3.3	3.5	99.3	327.4	523.3	244.5	148.1	2.7	3.7	**100.0**
NS LSTM (baseline)	200.0	93.8	154.0	158.1	5.1	1.7	**100.0**	225.8	280.0	208.0	138.2	3.0	2.4	**100.0**
VATES (proposed)	**96.6**	**85.1**	**109.8**	41.5	**1.7**	**1.1**	**100.0**	**212.1**	**264.2**	**170.8**	82.2	**2.5**	**2.0**	**100.0**

TABLE 3. Performance comparison on real data. NaN represents instances wherein the LLR test failed to detect crowding. The unit of MAE is the number of GPS logs, and the units of MASTE and MAETE are hours.

Model	Rugby World Cup Final						J1 League Final Section					
	MAE↓				MASTE↓	MAETE↓	MAE↓				MASTE↓	MAETE↓
	A	S	R	N			A	S	R	N		
EHA	228.0	665.5	103.0	33.8	5.0	2.0	361.4	620.0	**56.5**	24.9	4.0	3.0
BPReg* [18]	173.4	596.0	82.5	18.0	**0.0**	2.0	194.3	649.0	238.0	8.2	1.0	2.0
CityOutlook [12]	306.8	706.5	102.0	50.6	7.0	4.0	311.4	642.5	148.5	22.0	2.0	4.0
NS LSTM (baseline)	433.6	453.5	61.8	11.4	1.0	2.0	254.1	605.5	272.0	9.5	**0.0**	3.0
VATES (proposed)	**66.6**	**253.5**	**26.2**	15.2	**0.0**	**1.0**	**189.9**	**233.0**	149.5	12.5	1.0	**1.0**

Model	Comic Market						Tokyo Motor Show					
	MAE↓				MASTE↓	MAETE↓	MAE↓				MASTE↓	MAETE↓
	A	S	R	N			A	S	R	N		
EHA	152.6	121.0	79.0	19.9	NaN	NaN	85.3	**33.4**	99.8	8.4	**2.0**	5.0
BPReg* [18]	54.0	**18.5**	40.3	11.4	**0.0**	1.0	181.0	198.4	138.8	9.4	NaN	NaN
CityOutlook [12]	91.0	165.0	67.0	26.3	1.0	1.0	174.7	359.8	76.8	12.9	**2.0**	4.0
NS LSTM (baseline)	51.8	101.5	46.7	8.4	**0.0**	1.0	126.3	122.0	61.3	9.5	8.0	**0.0**
VATES (proposed)	**34.4**	71.5	**31.0**	11.5	**0.0**	**0.0**	**64.7**	64.6	**35.2**	5.4	3.0	1.0

Table 3 further confirms the promising forecasting performance of VATES across all error metrics in real events. From the table, EHA exhibited acceptable performance in the R state of the J1 League Final Section and the S state of Tokyo Motor Shows. However, this result was attributed to the coincidental pattern present in the training data, which is further discussed in the qualitative evaluation in Section V-D. BPReg* achieved the best MAE results for the S state in the Comic Market, while simultaneously presenting worse results across other metrics. CityOutlook demonstrated superior results in MASTE for the Tokyo Motor Show; however, it underperformed in the metrics for all other events. Conversely, in the case of VATES, we observed a consistent improvement in MAE for the A state and either the best or second-best results for MAE in the S state, MASTE, and MASTE across all events, while simultaneously producing the smallest errors in MAE for the R state in the three events except for the J1 League Final Section. We further discuss the fact that there were cases that VATES underperformed the baselines in Section V-D.

VATES did not provide an accurate forecasting of crowd density in the N state, which is out of our research scope. The prediction of non-crowded states can be performed by other models (e.g., EHA) given the accurately predicted crowding start and end by VATES.

Note that, in VATES, the overall improvement in MASTE was smaller than the improvement in MAETE. This could be because state-of-the-art methods, such as CityOutlook, were specialized to forecast the start of crowding.

We further compared our model with the survival analysis (SA) [10]-based approaches. Since the SA-based method predicts the end time of the crowding, we evaluate the performance of these methods only in MAETE. Table 4 and Table 5 list the forecasting performance.

TABLE 4. Performance comparison between SA [10]-based methods and VATES on SYNTHETIC DATA in MAETE↓ [h].

Model	forecast ahead	sports-type	exhibition-type
SA+BPReg	**1 week**	2.9	2.9
SA+CityOutlook	**1 week**	1.9	2.2
SA+VATES	**1 week**	1.6	2.7
VATES (proposed)	**1 week**	**1.1**	**2.0**
SA [10]	5 hours	0.9	1.2

From the tables, the original SA (listed in the bottommost), which can only predict the end time of crowding 5 h ahead,

TABLE 5. Performance comparison between SA [10]-based methods and VATES on REAL DATA in MAETE↓ [h].

Model	forecast ahead	Rugby	J1 Final
SA + BPReg	**1 week**	3.6	1.8
SA + CityOutlook	**1 week**	3.0	2.0
SA + VATES	**1 week**	3.0	1.6
VATES (proposed)	**1 week**	**1.0**	**1.0**
SA [10]	5 hours	1.4	1.0
Model	forecast ahead	**Comiket**	**Motor**
SA + BPReg	**1 week**	3.6	2.0
SA + CityOutlook	**1 week**	3.6	2.0
SA + VATES	**1 week**	3.4	**1.0**
VATES (proposed)	**1 week**	**0.0**	**1.0**
SA [10]	5 hours	1.4	0.6

performed better than VATES in the results with the synthetic data, and in the Tokyo Motor Show. However, none of the extensions (i.e., SA + BPReg, CityOutlook, VATES, which are listed above VATES) for one-week-ahead forecasting achieved better performance than VATES. This is because the input feature of these extensions is the forecasted crowd densities, which consist of forecasting errors.

D. QUALITATIVE PERFORMANCE EVALUATION

To understand how VATES behaves in real-world events, we visualized the crowd densities predicted by VATES and the baselines on the event date in Fig. 6.

In Fig. 6(a), BPReg*, which demonstrated superior performance in MAE for the A state and MASTE, simply predicted a density increase at approximately 16:00, followed by a predicted immediate decrease. Fig. 6(b) shows a similar trend, with BPReg* (and NS LSTM) resulting in an underestimation of density increase and early decrease. The output of EHA, which showed the minimum MAE for the R state in Section V-C, illustrated a gradual increase, consequently leading to an accurate density decline. However, VATES accurately captured the increase, sustain, and decrease in the density of both sports-type events.

In Fig. 6(c), EHA and CityOutlook predicted an increase until 9:00 but overlooked the secondary A state within the 9:00 ∼ 11:00 window. Contrarily, BPReg*, NS LSTM, and VATES identified an increase twice in the double attack; NS LSTM accurately captured the first A state, whereas VATES outperformed the other models in predicting the second. However, an unexpected density reduction occurred at 9:00, despite the S state[5]. BPReg* minimized the preceding density increase and consequently provided the most accurate prediction. Although VATES maintained the density and overlooked the sudden decline, it subsequently demonstrated quantitatively and qualitatively superior forecasting.

[5]This may be likely because of several factors such as loss of GPS signal from stationary users and instances of phone shutdowns for battery conservation. However, the crowd density at Comic Market, known for its pre-opening queues, was presumably maintained.

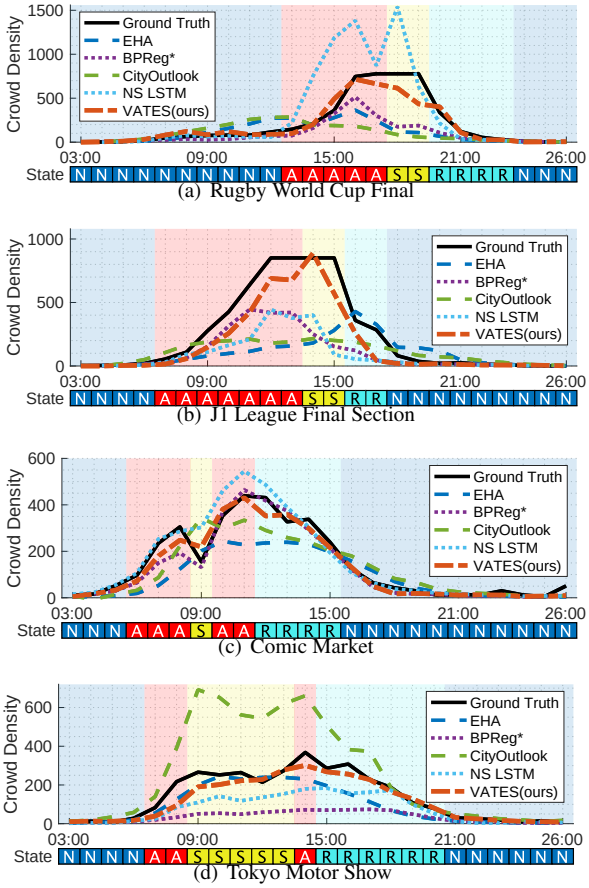


FIGURE 6. Visualization of forecasting results.

Finally, in Fig. 6(d), EHA and VATES methods demonstrated competitive forecasting for the first A and S states. Subsequently, EHA decreased the density, reconfirming its results as coincident, whereas VATES identified a second A state. These visualizations reveals VATES' promising performance in the LCE forecast.

E. CROWDING SYNTHESIS WITH MANIPULATED ENVELOPES

To further highlight the sensitivity and synthesizability of VATES to crowd states, we examined the predicted density transition under external envelope manipulation. After training, the model structure was adjusted to consider the envelope attributes (Eq. (7)) and state transitions (Eq. (12)) as external inputs. The manipulated envelopes were then fed into the model input.

Fig. 7 depicts the synthesized crowd density transitions. As shown in Fig. 7(a), providing an N input (Type 2) instead of an A input (Type 1) results in non-increased patterns of density. The result indicates that a reasonable crowd density increase has been modeled for the change from N to A, i.e., the start time of crowding. By feeding the model with S (Type 2) instead of A (Type 1), as shown in Fig. 7(b), the density temporarily stopped increasing during the sustained period.

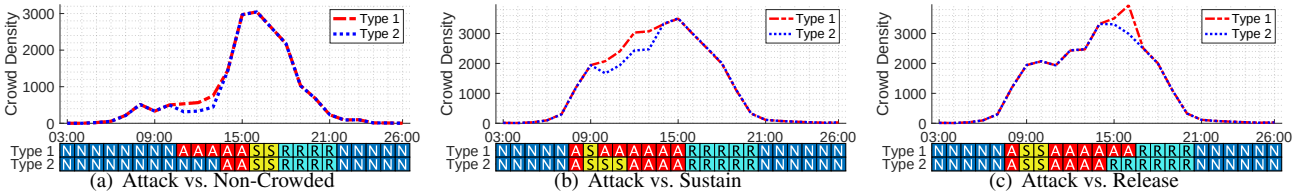


FIGURE 7. Result of crowding synthesis. The bottom shows the state transitions of the manipulated envelopes, and the top shows the synthesized density curves.

Furthermore, as demonstrated in Fig. 7(c), providing R (Type 2) instead of A (Type 1) triggered the model to begin reducing the density, in line with the state transition. From these results, we confirmed that the VATES was sensitive to the given envelope, indicating that VATES acquired the synthesizability of the heterogeneous crowd density transition.

F. ABLATION STUDY

VATES is characterized by (1) the envelope depiction (Section IV-C) and (2) state sequence labeling (Section IV-D). To analyze the importance of each contribution, we assessed the forecasting performance of the real-world four events using the following variants of VATES: (1) without envelope depiction, (2) without state sequence labeling, and (3) without both (akin to NS LSTM).

Table 6 summarizes the substantial contribution of the envelope depiction towards performance enhancement, while the state sequence labeling appeared to weaken performance compared to NS LSTM. Contrarily, integrating both elements, as in VATES, produces the best or near-best performance. However, the envelope illustration in Comic Market did not substantially contribute to MAE reduction during the Sustain and Release states, as it did in other events. This is probably owing to inaccurate crowding envelope prediction caused by a sudden decline during the Sustain state. Owing to the successful state sequence classification, VATES maintained a stable performance throughout this event.

VI. DISCUSSION

A. FINDINGS FROM THE EXPERIMENTAL RESULTS.

As discussed in Section II, existing crowding forecast and TSF methods can be divided into two groups: those that assume temporal autocorrelation [4], [8], [37], which is not assumable in the one week ahead of forecasting, and those that adopt side information as covariates [18], [46]. VATES belongs to those that use side information, and we compared VATES with existing side-information-based methods including the state-of-the-art method CityOutlook [46]. The synthetic data contained only 1% of events in 180 days of data, meaning that there were only one or two events in six months. The experimental results showed that even using the hours and purposes of the event, it was still not possible for the existing side-information-based methods to capture heterogeneous LCEs. However, modeling to capture the shape in the VATES framework enabled accurately predicting the LCEs.

The VATES was also compared to a recently proposed SA-based method [10] for predicting the duration of the crowded events. The experimental results showed that although the VATES (forecasting one week ahead) was partially inferior in performance to the original SA-based method (forecasting 5-h ahead), the VATES was consistently superior to the extension of the SA with [18], [46] for forecasting one week ahead. Surprisingly, there were also the cases where the VATES performed better or similarly to the SA-based 5-hour-ahead forecast.

Furthermore, we showed the crowding synthesis results in Section V-E. The results indicate that the VATES can simulate the LCE for fictional events by assuming a certain state transition. This simulation could be critical for anticipating contingencies under crowded events. Based on these results, we believe that the VATES is durable enough for one-week-ahead crowding forecasting in the real world.

B. APPLICABILITY OF VATES TO OTHER EVENT PURPOSES.

We believe that the applicability of VATES is not limited to sports games and exhibition events, but extends to other event purposes (e.g., fireworks displays, festivals). As mentioned in Section I, all crowded events have phases where the crowd density increases from the start of crowding and decreases towards the end of crowding, regardless of the purpose of the event. VATES captures such starts and ends through the shape modeling, thus it should be possible to forecast other events. Our future work will address other types of event purposes.

C. LIMITATIONS AND EXCEPTIONS.

We are aware that VATES may have a limitation, that is, the state transition is performed *manually* for the *predefined* event purposes. This may make the state transition infeasible in the following two cases: (1) when multiple event purposes are mixed in a single event, and (2) when the event is unprecedented. To address this issue, it may be beneficial to embed events into feature vectors. Events often have descriptions available in advance. By using pretrained large language models (LLMs) such as GPT-4 [47], a vector representation of events could be obtained from the description. Therefore, it may be possible to find precedented events that are close to the events with mixed or unprecedented purposes. We can also cluster the events by using the vector representation, thus it may be possible to automatically extract shapes and states from the average of the crowd density transitions for events that belong to the same cluster. We plan to address these issues

TABLE 6. Predictive performance comparison for ablation study at four events. The unit of MAE is the number of GPS logs, and the units of MASTE and MAETE are hours.

Model	Rugby World Cup Final						J1 League Final Section					
	MAE↓				MASTE↓	MAETE↓	MAE↓				MASTE↓	MAETE↓
	A	S	R	N			A	S	R	N		
VATES (proposed)	**66.6**	253.5	**26.2**	15.2	**0.0**	**1.0**	**189.9**	**233.0**	**149.5**	12.5	1.0	**1.0**
w/o Envelope Depiction	920.6	1088.5	107.8	29.6	2.0	3.0	248.4	2325.5	348.0	7.5	**0.0**	**1.0**
w/o State Sequence Labeling	145.4	**114.0**	61.3	21.2	1.0	1.0	369.6	573.0	160.0	11.5	4.0	1.0
w/o both (= NS LSTM)	433.6	453.5	61.8	11.4	1.0	2.0	254.1	605.5	272.0	9.5	**0.0**	3.0

Model	Comic Market						Tokyo Motor Show					
	MAE↓				MASTE↓	MAETE↓	MAE↓				MASTE↓	MAETE↓
	A	S	R	N			A	S	R	N		
VATES (proposed)	34.4	71.5	31.0	11.5	**0.0**	**0.0**	**64.7**	64.6	**35.2**	5.4	**3.0**	1.0
w/o Envelope Depiction	**25.0**	**38.0**	**21.0**	12.9	**0.0**	1.0	131.3	100.6	76.2	7.3	9.0	2.0
w/o State Sequence Labeling	44.6	154.5	87.3	11.9	**0.0**	2.0	104.3	**62.4**	94.3	6.5	4.0	1.0
w/o both (= NS LSTM)	51.8	101.5	46.7	8.4	**0.0**	1.0	126.3	122.0	61.3	9.5	8.0	**0.0**

in future work.

There are exceptional events that VATES cannot forecast; VATES assumes that the event is publicly announced in advance. However, VATES cannot predict sudden crowding caused by events that are not scheduled in advance, such as flash mobs, spontaneous protests, or vigils for recent events. Although prior work [48] has shown that the SNS posts increase prior to such social events, which may suggest that the number of SNS posts can be used as an indicator of future crowd gatherings, it may be still difficult to forecast the LCE in such events with many uncertainties.

D. REPRODUCIBILITY OF VATES IN TERMS OF DATA.

GPS mobility log. Using GPS logs to capture crowd densities is not a problem in terms of reproducibility. In recent years, various applications and services have been logging the user's GPS-based locations, and various studies have been conducted using the mobility logs (e.g., "Konzatsu-Tokei (R)" from ZENRIN DataCom Co., Ltd. [4], a mobile application from LY Corporation [49], dataset from Tencent [50]).

Transit search log. We obtained people's scheduled visits from the transit search released by LY corporation; however, the other records such as route searching history on map applications (e.g., Google Maps, Yahoo! Map, ZENRIN Map, Japan Transit Planner, NAVITIME) or logs on travel reservation applications (e.g., Booking.com, Travelko) can also be used as the scheduled patterns.

VII. CONCLUSION

We have presented the VATES to forecast the LCE one week in advance, which no work had realized to date. Inspired by acoustic synthesis, we discussed the benefits of learning the shape of crowd density transition for forecasting the heterogeneous LCE. Experimental results using synthetic and real data validated the efficacy of our models. Compared with the state-of-the-art methods, the VATES showed a 24.3% performance improvement in MAE when predicting crowd density transitions during crowding, and 6.6% and 26.1% performance improvement in MASTE and MAETE when predicting the start and end times of crowding, respectively. We also confirmed the feasibility of synthesizing crowd densities by externally controlling the state transition, which indicated that the VATES could simulate fictional events. These results suggest that our method can enhance the safety and mobility of individuals in urban environments, thereby contributing to smarter city management and improving the quality of life for urban populations.

Future work will address other event purposes, such as firework displays and festivals, by tailoring state transitions for such events. We plan to investigate the extensive applicability of our model to numerous real-world events, including those that occurred after the COVID-19 epidemic. Automatic extraction of event purposes, which are now predefined by humans, should also be considered. Because descriptions written by event organizers are often public, we plan to leverage language embeddings generated by LLMs [51].

ACKNOWLEDGMENT

We would like to thank Editage (www.editage.jp) for English language editing.

REFERENCES

[1] Junbo Zhang et al. DNN-based prediction model for spatio-temporal data. In *Proc. of SIGSPATIAL*, 2016.
[2] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proc. of AAAI*, 2017.
[3] Ziqian Lin et al. DeepSTN+: Context-aware spatial-temporal neural network for crowd flow prediction in metropolis. In *Proc. of AAAI*, 2019.
[4] Renhe Jiang et al. DeepUrbanEvent: A system for predicting citywide crowd dynamics at big events. In *Proc. of SIGKDD*, 2019.
[5] Renhe Jiang et al. DL-Traff: Survey and benchmark of deep learning models for urban traffic prediction. In *Proc. of CIKM*, 2021.
[6] Chung Park et al. PASTA: Parallel spatio-temporal attention with spatial auto-correlation gating for fine-grained crowd flow prediction. In *Proc. of PAKDD*, 2022.
[7] Zhaonan Wang et al. Event-aware multimodal mobility nowcasting. In *Proc. of AAAI*, 2022.
[8] Renhe Jiang et al. Learning social meta-knowledge for nowcasting human mobility in disaster. In *Proc. of TheWebConf*, 2023.

[9] Amin Vahedian et al. Predicting urban dispersal events: A two-stage framework through deep survival analysis on mobility data. In *Proc. of AAAI*, 2019.
[10] Amin Vahedian Khezerlou et al. DILSA+: Predicting urban dispersal events through deep survival analysis with enhanced urban features. *ACM TIST*, 12(4):1–25, 2021.
[11] Tatsuya Konishi et al. CityProphet: City-scale irregularity prediction using transit app logs. In *Proc. of UbiComp*, 2016.
[12] Soto Anno et al. Cityoutlook: Early crowd dynamics forecast towards irregular events detection with synthetically unbiased regression. In *Proc. of SIGSPATIAL*, 2021.
[13] F Leccese et al. The bowed string instruments: acoustic characterization of unique pieces from the italian lutherie. In *IOP Conference Series: Materials Science and Engineering*, 2018.
[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.
[15] Yu Zheng et al. Diagnosing new york city's noises with ubiquitous data. In *Proc. of UbiComp*, 2014.
[16] Massimiliano Luca et al. A survey on deep learning for human mobility. *ACM CSUR*, 55(1), 2021.
[17] Minh X Hoang et al. FCCF: forecasting citywide crowd flows based on big data. In *Proc. of SIGSPATIAL*, 2016.
[18] Masamichi Shimosaka et al. Forecasting urban dynamics with mobility logs by bilinear poisson regression. In *Proc. of UbiComp*, 2015.
[19] Yuta Hayakawa et al. Simultaneous multiple poi population pattern analysis system with hdp mixture regression. In *Proc. of PAKDD*, 2021.
[20] Masamichi Shimosaka et al. Spatiality preservable factored poisson regression for large-scale fine-grained gps-based population analysis. In *Proc. of AAAI*, 2019.
[21] Soto Anno et al. Supervised-CityProphet: Towards accurate anomalous crowd prediction. In *Proc. of SIGSPATIAL*, 2020.
[22] Liang Zhao. Event prediction in the big data era: A systematic survey. *ACM CSUR*, 54(5):1–37, 2021.
[23] Dawei Wang et al. Towards long-lead forecasting of extreme flood events: a data mining framework for precipitation cluster precursors identification. In *Proc. of SIGKDD*, 2013.
[24] Aleksandr Simma. *Modeling events in time using cascades of Poisson processes*. University of California, Berkeley, 2010.
[25] Alexander Ihler et al. Adaptive event detection with time-varying poisson processes. In *Proc. of SIGKDD*, 2006.
[26] Nan Du et al. Recurrent marked temporal point processes: Embedding event history to vector. In *Proc. of SIGKDD*, 2016.
[27] Maya Okawa et al. Deep mixture point processes: Spatio-temporal event prediction with rich contextual information. In *Proc. of SIGKDD*, 2019.
[28] Kaiqun Fu et al. TITAN: A spatiotemporal feature learning framework for traffic incident duration prediction. In *Proc. of SIGSPATIAL*, 2019.
[29] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.
[30] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.
[31] David S Stoffer and Hernando Ombao. Special issue on time series analysis in the biological sciences, 2012.
[32] Torben G Andersen et al. Volatility forecasting, 2005.
[33] Junyoung Chung et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
[34] Alex Krizhevsky et al. ImageNet classification with deep convolutional neural networks. *Proc. in NIPS*, 25, 2012.
[35] Shaojie Bai et al. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
[36] Du Tran et al. A closer look at spatiotemporal convolutions for action recognition. In *Proc. of CVPR*, 2018.
[37] Tyler Wilson et al. DeepGPD: A deep learning approach for modeling geospatio-temporal extreme events. In *Proc. of AAAI*, 2022.
[38] Giovanni De Poli. A tutorial on digital sound synthesis techniques. *Computer Music Journal*, 7(4):8–26, 1983.
[39] Diemo Schwarz and Xavier Rodet. Spectral envelope estimation, representation, and morphing for sound analysis, transformation, and synthesis. In *Proc. of ICMC*, 1999.
[40] D Brandon Lloyd et al. Sound synthesis for impact sounds in video games. In *Symposium on Interactive 3D Graphics and Games*, pages 55–62, 2011.
[41] Rainer Kelz et al. Deep polyphonic adsr piano note transcription. In *Proc. of ICASSP*, 2019.
[42] Daniel B. Neill. Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 2009.
[43] Xun Zhou et al. A traffic flow approach to early detection of gathering events. In *Proc. of SIGSPATIAL*, 2016.
[44] Matei Zaharia et al. Spark: Cluster computing with working sets. *HotCloud*, 2010.
[45] Jason (Jinquan) Dai et al. BigDL: A distributed deep learning framework for big data. In *Proc. of the ACM Symposium on Cloud Computing*, 2019.
[46] Soto Anno et al. CityOutlook+: Early crowd dynamics forecast through unbiased regression with importance-based synthetic oversampling. *IEEE Pervasive Computing*, 22(4):26–34, 2023.
[47] Josh Achiam et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
[48] Liang Zhao et al. Spatiotemporal event forecasting in social media. In *Proc. of SIAM SDM*, 2015.
[49] Takahiro Yabe et al. A framework for evacuation hotspot detection after large scale disasters using location data from smartphones: Case study of kumamoto earthquake. In *Proc. of SIGSPACIAL*, 2016.
[50] Tong Xia and Yong Li. Revealing urban dynamics by learning online and offline behaviours together. *Proc. of IMWUT*, 2019.
[51] Sébastien Bubeck et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

SOTO ANNO received his BE and ME from the Tokyo Institute of Technology in 2020, 2022. He is currently a Ph.D. candidate student at the Tokyo Institute of Technology. His research focuses on urban computing. Contact him at anno@miubiq.cs.titech.ac.jp.

KOTA TSUBOUCHI received the PhD degree from the University of Tokyo, Japan, in 2010. Until March 2012, he did research about on-demand traffic systems with the University of Tokyo. Since April 2012, he has been a data scientist and a senior researcher with Yahoo JAPAN Research. His research interest focuses on data analysis including human activity logs, such as location information, search logs, shopping history, and sensor data. Contact him at ktsubouc@yahoo-corp.jp.

MASAMICHI SHIMOSAKA received his BE, ME, and Ph.D. from the University of Tokyo in 2001, 2003, and 2006, respectively. He joined Tokyo Institute of Technology as an associate professor in July 2015. Prior to joining Tokyo Institute of Technology, he was a faculty member at the University of Tokyo from 2006 to 2015. While his Ph.D. candidate, he was funded by the Japan Society for the Promotion Science as a research fellow. His research interests include machine intelligence and ubiquitous computing. He is a member of the ACM and IEEE. Contact him at simosaka@miubiq.cs.titech.ac.jp.