STICAP: Spatio-temporal Interactive Attention for Citywide Crowd Activity Prediction

HUIQUN HUANG, SUINING HE, XI YANG, and MAHAN TABATABAIE, University of Connecticut, USA

Accurate citywide crowd activity prediction (CAP) can enable proactive crowd mobility management and timely responses to urban events, which has become increasingly important for a myriad of smart city planning and management purposes. However, complex correlations across the crowd activities, spatial and temporal urban environment features and their *interactive* dependencies, and relevant external factors (e.g., weather conditions) make it highly challenging to predict crowd activities accurately in terms of different venue categories (for instance, venues related to dining, services, and residence) and varying degrees (e.g., daytime and nighttime).

To address the above concerns, we propose STICAP, a citywide spatio-temporal interactive crowd activity prediction approach. In particular, STICAP takes in the location-based social network check-in data (e.g., from Foursquare/Gowalla) as the model inputs and forecasts the crowd activity within each time step for each venue category. Furthermore, we have integrated multiple levels of temporal discretization to interactively capture the relations with historical data. Then, three parallel *Residual Spatial Attention Networks* (RSAN) in the *Spatial Attention Component* exploit the hourly, daily, and weekly spatial features of crowd activities, which are further fused and processed by the *Temporal Attention Component* for *interactive CAP*. Along with other external factors such as weather conditions and holidays, STICAP adaptively and accurately forecasts the final crowd activities per venue category, enabling potential activity recommendation and other smart city applications. Extensive experimental studies based on three different real-world crowd activity datasets have demonstrated that our proposed STICAP outperforms the baseline and state-of-the-art algorithms in CAP accuracy, with an average error reduction of 35.02%.

CCS Concepts: • **Information systems → Spatial-temporal systems;**

Additional Key Words and Phrases: Crowd activity prediction, spatial attention, temporal attention, external factors

ACM Reference format:
Huiqun Huang, Suining He, Xi Yang, and Mahan Tabatabaie. 2024. STICAP: Spatio-temporal Interactive Attention for Citywide Crowd Activity Prediction. *ACM Trans. Spatial Algorithms Syst.* 10, 1, Article 3 (January 2024), 22 pages.
https://doi.org/10.1145/3603375

H. Huang, S. He, and X. Yang, equal contributions.

This project was supported, in part, by the National Science Foundation (NSF) under Grant 2118102.

Authors' address: H. Huang, S. He (Corresponding author), X. Yang, and M. Tabatabaie, University of Connecticut, Department of Computer Science & Engineering, 371 Fairfield Way Unit 4155, Storrs, Connecticut 06269; e-mails: {huiqun.huang, suining.he, xi.yang, mahan.tabatabaie}@uconn.edu.

2374-0353/2024/01-ART3 $15.00
https://doi.org/10.1145/3603375

1 INTRODUCTION

The proliferation of mobile devices and emergence of location-based social network services have provided unprecedented opportunities of characterizing and understanding the citywide *crowd mobility patterns*. Further prediction of the *crowd participation levels* with respect to different venue categories (say, education, dining, or public transportation), namely, ***crowd activity prediction* (CAP)**, has started to gain much attention recently due to the significant values for smart city management and planning. Through the monitored and forecasted crowd activity participation within different venue categories, the city planners can proactively determine the allocation of essential civil infrastructures and mobility-related resources.

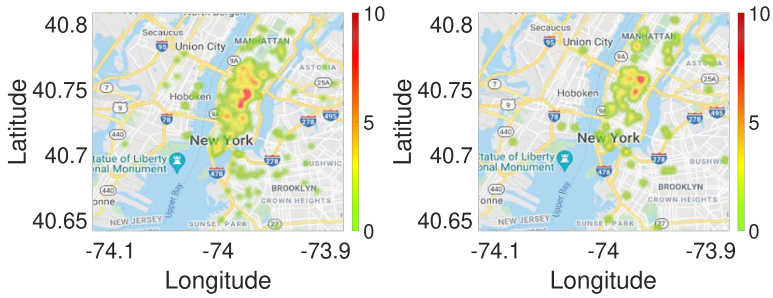
Despite a few prior efforts [13, 14, 18, 28, 29, 33–35], there remain the following major challenges in realizing accurate CAP. The spatial distributions of crowd activities of the same category in different regions are highly correlated with complex *interactions*, while those of different categories might interact with each other in varying spatial and temporal degrees. As illustrated in Figure 1, the dining (food) activities of crowds in New York City (NYC) during 04/02/2012–08/22/2012 demonstrated the similar frequencies among the surrounding regions, while the crowd activities of shop & service categories may be shown to be largely concentrated in the Manhattan area. Furthermore, crowd activities of different venue categories may often demonstrate the short-/long-term patterns, while the crowd routines might vary across hours, days, and weeks. For instance, we can see the daily routines (05/14/2012–05/16/2012) and weekly trend (04/02/2012–08/22/2012) of crowd activities in Figure 2.

To address the above challenges, we proposed STICAP, a novel **S**patio-**T**emporal **I**nteractive **C**rowd **A**ctivity **P**rediction approach. As illustrated in Figure 3, we aim to leverage the time- and location-tagged location-based social network check-in data from Foursquare/Gowalla [9] to model people's activity status in different venue categories. To predict crowd activity, we select several key features through extensive analysis upon the real-world check-in, weather, and other context-related datasets. We have designed a spatio-temporal residual attention-based model for CAP. The proposed model interactively differentiates the spatial and temporal characteristics of the check-in data and accounts for the impacts of external factors of temperature, wind speed, weather context, weekdays, and holidays on the CAP performance.

To summarize, we have made the following three major contributions:

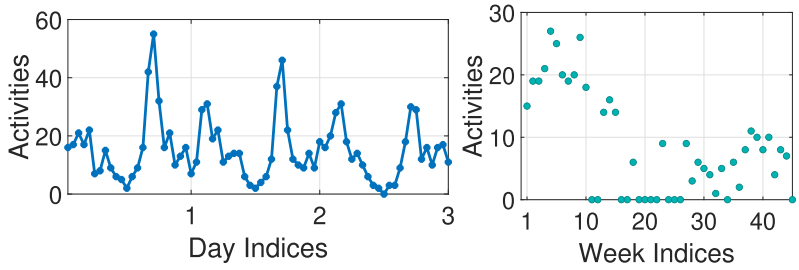
(1) **Comprehensive Interaction Analysis with Crowd Activity Data.** We have conducted extensive and comprehensive crowd activity data analysis based on real-world Foursquare [2] and Gowalla [1] check-in data with more than 500k records in three metropolitan cities (New York City, Los Angeles, and Tokyo) and identified the important spatial and temporal characteristics and interactions of the crowd activities with respect to different venue categories and the correlations with the external factors such as the weather conditions and their impacts. These important insights will serve as the foundation of our core STICAP designs.

(2) **Spatio-temporal Interactive Residual Attention Module for Crowd Activity Prediction.** To further capture the spatio-temporal interactions, we have designed the *Spatial Attention Component*, *Temporal Attention Component*, and *External Component* within our proposed STICAP. We have designed a novel **Residual Spatial Attention Network (RSAN)** in the *Spatial Attention Component*, along with interactive integration of the *Temporal Attention Component* to capture and differentiate the spatio-temporal characteristics of crowd activities, resulting in high accuracy in CAP problem. With the RSAN, the Spatial Attention Component accounts for and differentiates the *spatial interactions* of the crowd activities from different regions within the short- and long-term range. This way, STICAP captures the



(a) Dining (Food). (b) Shops & Services.

Fig. 1. Illustration of crowd activity heatmaps of food and shop & service categories of NYC in 04/12/2012.



(a) Daily trend. (b) Weekly trend.

Fig. 2. (a) Transportation-related crowd activities in NYC (05/14/2012–05/16/2012); (b) food-related weekly crowd activities in NYC at 5pm on every Friday (04/02/2012–08/22/2012).

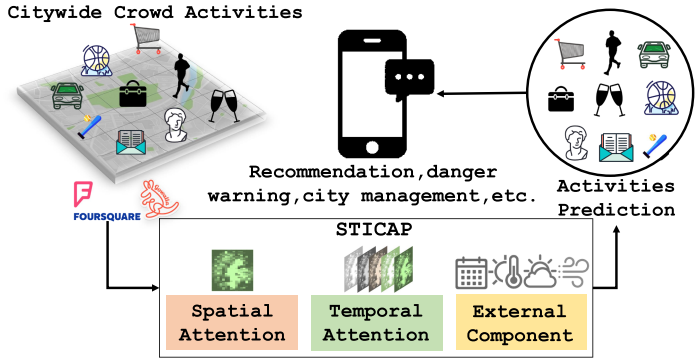


Fig. 3. Illustration of the citywide crowd activity participation monitoring and prediction.

dynamically varying contributions of different regions at different time steps. Specifically, for the RSAN, we have designed a *trunk branch* and a *mask branch* to, respectively, process and select the most important spatial features across different regions and time steps. The Temporal Attention Component further captures the varying *temporal interactions* of crowd activities from different time steps. With the complex spatio-temporal interactions captured by the spatio-temporal attention, STICAP adapts to the complex crowd activity, leading to accurate CAP results.

(3) **Extensive Experimental Evaluations with Real-world Datasets.** We have conducted extensive experimental evaluations of STICAP based on crowd activity datasets from three metropolitan cities (two in North America and one in East Asia). Specifically, we have evaluated our STICAP and other CAP approaches with two Foursquare datasets, including 153,610 check-in records in **New York City (NYC)** and 330,208 check-in records in Tokyo, Japan, as well as a Gowalla dataset with 51,305 check-in records in **Los Angeles (LA)**. Compared to the baselines and the state-of-the-arts (such as CHAT [15], ST-Norm [10], FDW [6], and DeepST [38]), our results show that STICAP demonstrates, on average, 35.02% improvements in the CAP accuracy.

The rest of the article is organized as follows: We first review the related work in Section 2, followed by data analysis, important concepts, and problem formulation in Section 3. The details of our STICAP framework will be introduced in Section 4, followed by the experimental evaluation in Section 5. We finally discuss the deployment of STICAP in Section 6 and conclude our work in Section 7.

2 RELATED WORK

We briefly overview the related work as follows:

—**Urban Mobility Applications.** With massive urban mobility data generated by ubiquitous devices [5], crowd mobility analytics have attracted extensive attention due to its significant social and business values in global market. On location-based social network platforms, it is viable to leverage users' historical activity data and provide personalized recommendation services, such as recommendation of **points-of-interest (POIs)**, group-oriented advertisement [36], activity for certain individuals or social groups [17, 23, 25, 39], and trip recommendation [31]. However, the crowd flow prediction has received wide attention, and the existing studies largely focus on applications such as the urban crowd momentum monitoring [4, 8, 11, 30], human trajectory prediction [12, 37], and abnormal event detection/prediction [16, 26].

However, forecasting *crowd participation levels in different activity venue categories*, i.e., crowd activity prediction or CAP, remains largely unexplored. To fill this gap towards a key enabler of predictive crowd distribution planning, our work focuses on designing a novel data-driven and spatio-temporal interactive residual attention learning-based approach to realize accurate CAP.

—**Crowd Mobility Learning.** The large-scale crowd activity data makes it highly challenging for applying the conventional statistical methods on crowd activity prediction. Therefore, many deep learning approaches have been widely adopted to unleash their computing power [20, 21].

The prior arts such as DeepST [38] consider the crowd flow in a city as image-like data with three different timeframes by partitioning the city map into grids and provide the deep learning model to capture the spatial and temporal features of such image-like data. However, the design of DeepST does not consider differentiating the dynamic contributions of different time steps, thus yielding low accuracy in the dynamic crowd activity prediction. Zhou et al. [40] employed an encoder-decoder framework, which excels in short-term prediction for multi-step citywide passenger demand prediction problem. Huang et al. [16] proposed the multi-head spatio-temporal attention mechanism to predict the occurrence of citywide abnormal events by considering the correlations among the historical citywide crowd flow movement and the occurrence of abnormal events.

Different from the above studies, we propose the interactive attention mechanism in citywide crowd activity prediction problem. With spatial-temporal attention mechanism, we consider the spatial dependencies in three levels of time ranges, temporal dependency of the check-in data, and the impacts of external factors in each time step, to capture the semantic information within the

Table 1. Overview of Various Datasets Adopted in STICAP's Designs

City	Check-ins		Weather	
NYC	User ID, venue ID, venue category ID, venue category, latitude, longitude, and timestamp.	153,610 check-ins, [40.55085247°*N*, 40.98833172°*N*], [73.6838252°*W*, 74.27476645°*W*].	Temperature, wind speed, and weather condition.	3,432 records, temperature: [12, 100] °*F*, wind speed: [3, 56] mph.
Tokyo		330,208 check-ins, [35.51018469°*N*, 35.86715042°*N*], [139.4708776°*E*, 139.9125931°*E*]		3,432 records, temperature: [28, 93] °*F*, wind speed: [1, 61] mph.
LA		51,305 check-ins, [33.6099916°*N*, 34.1813999°*N*], [117.5323391°*E*, 118.4984708°*E*]		3,624 records, temperature: [33, 109] °*F*, wind speed: [3, 32] mph.

check-in data. We note that our STICAP differs from the prior studies [38] in the following two important aspects: *(a)* We have designed the trunk branch structure to capture the deep spatial features across different regions and time steps and leverage the mask branch structure to quantify their weights. This way, STICAP enables more *interactive* and *flexible* characterization of the spatio-temporal crowd activities. *(b)* We have designed temporal attention to measure and characterize the pair-wise correlations across different time steps, which overcome the limits to certain time periods defined in fixed time spans [38]. Our experimental studies further validate the importance of the aforementioned designs.

The attention mechanisms have been widely studied for time-series prediction [24], video processing [7], sequence classification [22], geo-sensory time series prediction [19], and passenger demand prediction [40]. Compared with the recent attention-based works for spatio-temporal data like GeoMAN [19], we note that our work focuses on modeling the *interactions* through temporal attention weights, integrating the close, medium, and distant historical records to differentiate their impact upon the predictions. Furthermore, while GeoMAN considers attention upon the geo-sensory time series prediction for individual sensors, our work takes into account more comprehensive spatial interaction knowledge, i.e., the spatial attentions upon the input grid-based heatmaps. Addressing the prior studies' limitations of prediction effective only for short-term periods, our proposed adaptive method in STICAP integrates multi-level spatial and temporal information with spatial-temporal attention modules with high accuracy.

3 DATASETS, CONCEPTS, & PROBLEM FORMULATION

This section outlines the various datasets studied in Section 3.1, introduces the important concepts and problem definitions of this study in Section 3.2, followed by the spatio-temporal data analysis in Section 3.3.

3.1 Datasets Studied

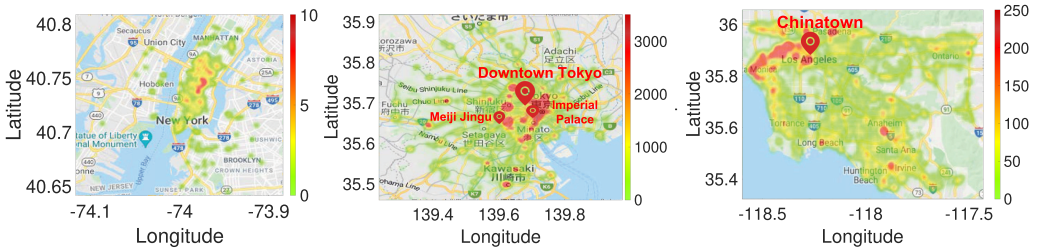
In this study, we use three types of datasets (summarized in Table 1) to model the crowd activities of NYC, Tokyo, and LA. The details of each dataset are introduced as follows

— **Crowd Activity Data.** In this study, Foursquare check-in data [2] during04/02/2012–08/22/2012 in NYC and Tokyo, and Gowalla check-in data [1] during 12/01/2009–04/30/2010 in LA are collected and utilized to represent the crowd activities. The crowd activity data of NYC include 153,610 check-in records from 1,083 distinct users. The data of Tokyo include 330,208 check-in records from 2,293 distinct users, while the data of LA include 51,305 check-in records. Each check-in record is composed of the user ID, venue category, latitude, longitude, and check-in timestamp. An example of Foursquare check-in record is shown in Table 2.

The venue category of the crowd activities can reflect the type of crowd activities relevant to the urban venue functions. This study divides the crowd activities into nine categories for NYC and

Table 2. A Location-based Social Network Check-in Example

User Id	Venue Category	Latitude	Longitude	UTC Timestamp
1541	Cosmetics Shop	35.70510109°N	139.61959°E	Tue, Apr 03, 18:17:18 +0000, 2012



(a) NYC. (b) Tokyo. (c) LA.

Fig. 4. Interactive crowd activity distribution heatmaps of (a) NYC, (b) Tokyo, and (c) LA.

Tokyo, and 10 categories for LA based on the venue functions in the Foursquare/Gowalla check-in datasets.

—**External Factors.** To analyze the influence of weather on crowd activity participation, we collect hourly weather temperature, wind speed, and weather contexts from the open-source weather data API [3] for NYC, Tokyo, and LA. Weather contexts are given by the phrases/words used for describing the weather status, such as *windy*, *cloudy*, and so on. The data are collected during 04/02/2012–08/22/2012 for NYC and Tokyo, and collected during 12/01/2009–04/30/2010 for LA.

To further study the interactive impact of the weekends and holidays on the occurrences of crowd activities, we categorize the crowd activity based on whether it is on weekdays/weekends and on holidays/non-holidays. Specifically, we use an indicator to denote weekdays as 1 and weekend as 0, and we use another indicator to denote holidays as 1 and non-holiday periods as 0.

3.2 Important Concepts & Problem Definition

We present the important concepts in this study as follows:

—**City Region.** In this study, we first build the smallest quadrilateral for NYC, Tokyo, and LA using all the distinct locations of the check-in data of each city, respectively. The details of the data are shown in Table 1. The quadrilateral is then partitioned into a $I \times J$ grid map based on the longitudes and latitudes. Every grid in the grid map is considered as a distinct region. The check-in heat maps using all the check-in data of NYC, Tokyo, and LA are shown in Figures 4(a), 4(b), and 4(c), respectively.

—**Problem Definition.** We present the problem statement of STICAP as follows: Let T be the number of time steps discretized within a day, V be the number of crowd activity categories. Given:

(1) *Near History Records*: which represent the L time steps of near history crowd activity heatmaps $\mathbf{G}_c \in \mathbb{R}^{L \times I \times J}$ during the time period $\{t-L, \ldots, t-1\}$;
(2) *Historical Daily Records*: which represent the L time steps of historical daily crowd activity heatmaps $\mathbf{G}_p \in \mathbb{R}^{L \times I \times J}$ during the time period $\{t-LT, \ldots, t-T\}$;
(3) *Historical Weekly Records*: the L time steps of historical weekly crowd activity heatmaps $\mathbf{G}_{\mathtt{tr}} \in \mathbb{R}^{L \times I \times J}$ during the time period $\{t-7LT, \ldots, t-7T\}$; and
(4) *External Factors*: which represent the d time steps of external factor vector, denoted as $\mathbf{E} \in \mathbb{R}^{L \times p}$, during the time period $\{t-L, \ldots, t-1\}$,

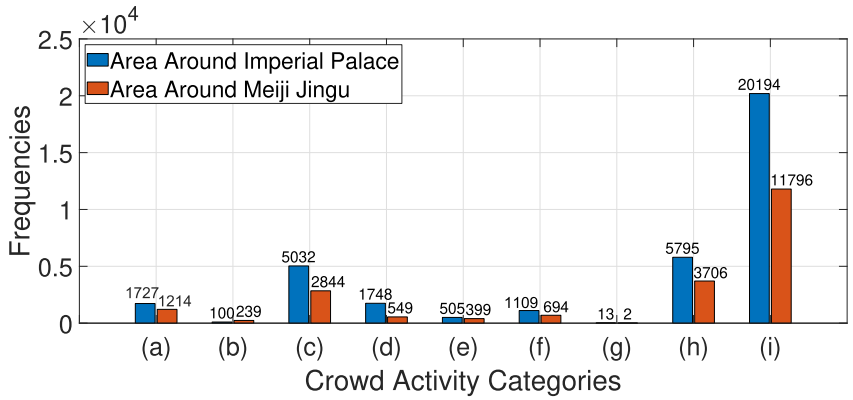


Fig. 5. Crowd activities of categories of *(a)* arts & entertainment, *(b)* education, *(c)* food, *(d)* nightlife spot, *(e)* outdoor, *(f)* professional, *(g)* residential, *(h)* shop & service, and *(i)* transportation in the areas around the Imperial Palace and Meiji Jingu of Tokyo during 04/02/2012–08/22/2012.

the *goal* of STICAP is to forecast the occurrences of the total V categories of crowd activities, denoted as $\hat{\mathbf{X}}_t \in \mathbb{R}^V$, in the future time step t. The problem of STICAP can be formally written as

$$\hat{\mathbf{X}}_t \sim \mathcal{F}(\mathbf{G}_c; \mathbf{G}_p; \mathbf{G}_{\mathtt{tr}}; \mathbf{E}). \tag{1}$$

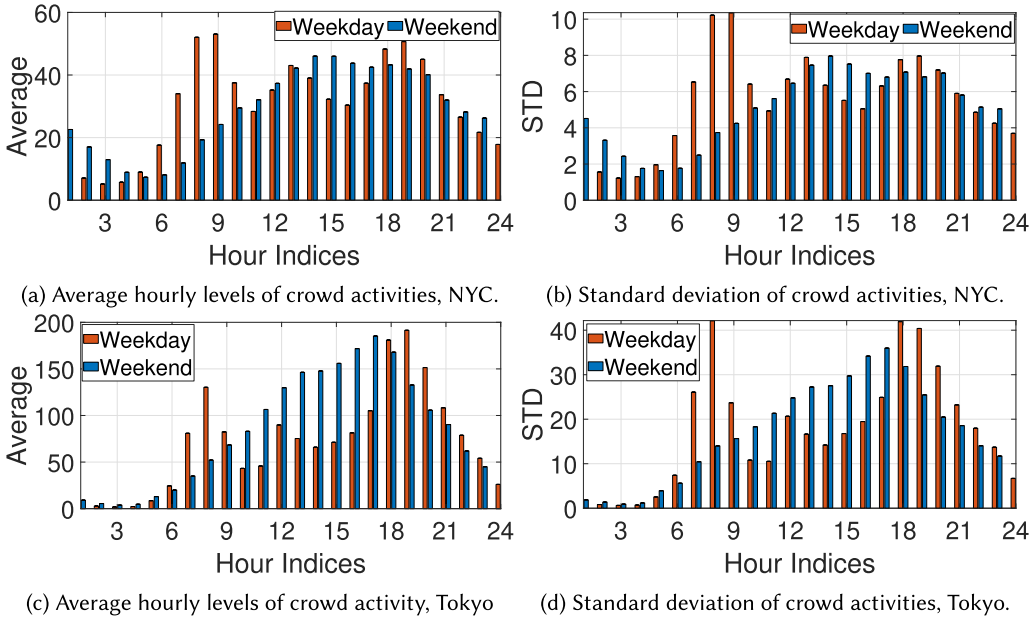
3.3 Data Analysis on Spatio-temporal Interaction

—**Spatial Interaction Analysis.** We note that different function zones within a city may shed an impact upon the spatial distributions and participation levels of various crowd activities. Figures 4(a), 4(b), and 4(c) first overview all the crowd activity heatmaps of NYC, Tokyo, and LA, respectively. The warmer color in the maps represents more check-ins. In Figure 4(a), we can observe that the major crowd activities come from Manhattan, NYC. while in Figures 4(b) and 4(c), we have observed that the crowds mainly distribute in the downtown areas of Tokyo and LA. The dense crowd activity distributions within or around the center of the cities with multiple different city functions imply the spatial interaction of the crowds with these complex city function zones, which need to be comprehensively captured for accurate CAP.

Furthermore, we can also observe the spatial *heterogeneous* interactions of crowd activities with the complex city function zones. Taking Tokyo as an example, in Figure 5, we can observe that the crowd activities of nine categories of Tokyo during 04/02/2012–08/22/2012 in the areas around the Imperial Palace and Meiji Jingu (as highlighted in Figure 4(b)) demonstrated a noticeable disparity in terms of measured levels, even if these two areas are both in downtown Tokyo. How to capture these heterogeneous interactions is challenging for accurate CAP.

—**Interactive Crowd Activities during Weekdays/Weekends/Holidays.** To further analyze the correlations between the occurrences of crowd activities on weekdays and weekends, we calculate the average occurrences of crowd activities in each hour of a day using both the crowd activities during weekdays and weekends. Specifically, the average occurrences of crowd activities in the lth hour of the weekday, denoted as $\mathbf{AVG}^l_{wd} \in \mathbb{R}$, and that of the weekend, denoted as $\mathbf{AVG}^l_{wk} \in \mathbb{R}$, are given by

$$\mathbf{AVG}^l_{wd} = \frac{\sum_{h=1}^{D_{wd}} Z^{h,l}_{wd}}{D_{wd}}, \quad \mathbf{AVG}^l_{wk} = \frac{\sum_{h=1}^{D_{wk}} Z^{h,l}_{wk}}{D_{wk}}, \tag{2}$$



(a) Average hourly levels of crowd activities, NYC.

(b) Standard deviation of crowd activities, NYC.

(c) Average hourly levels of crowd activity, Tokyo

(d) Standard deviation of crowd activities, Tokyo.

Fig. 6. The average levels and standard deviations of hourly crowd activities during weekdays/weekends of NYC and Tokyo.

where D_{wd} and D_{wk} denote the total days of check-ins during weekdays and weekends, respectively, $Z_{wd}^{h,l}$ denotes the total crowd activities on the lth hour of day h in D_{wd}, and $Z_{wk}^{h,l}$ denotes the total check-ins on the lth hour of day h in D_{wk}, $l \in \{0, \ldots, 23\}$. Similarly, we also find the **standard deviations (STD)** of the occurrences of the crowd activities during the weekends and weekdays, respectively.

As shown in Figure 6(a) and Figure 6(c), there are generally three peaks of crowd activity on a weekday for NYC and Tokyo. The peaks are around the periods of 6am–9am, 11am–1pm, and 4pm–8pm for NYC and Tokyo. Crowd activities of NYC and Tokyo on weekends share the similar patterns, i.e., the occurrences of crowd activities mostly increase from around 5am–1pm and decrease from around 1pm to around mid-night. Extracting the short/long-term temporal patterns of crowd activities and differentiating the temporal interactions of crowd activities from different time steps may enhance the accuracy of the CAP. Furthermore, we can observe larger variations during late afternoon and early evening, particularly for the weekend (Figures 6(b) and 6(d)), which imply the potentially interactive behaviors likely due to the various life-related or recreational activities.

We also show the influence of the holiday events upon the crowd activities. As shown in Figures 7(a) and 7(b), we, respectively, show the levels of crowd activities related to transportation category in NYC in terms of historical hourly average (blue) and the ones (orange) with the impacts of Independence Day (07/04/2012, Wednesday) and Thanksgiving Day (11/22/2012, Thursday), which are public holidays in the U.S.. This motivates us to take into account these holiday events for our CAP.

—**Impacts of Weather Conditions upon Crowd Activities.** We further show the impacts of the weather conditions upon the crowd activities by analyzing the difference of crowd activities given different weather conditions. In Figures 8(a) and 8(b), we, respectively, demonstrate the levels of crowd activities related to transportation category in terms of historical hourly average (blue)

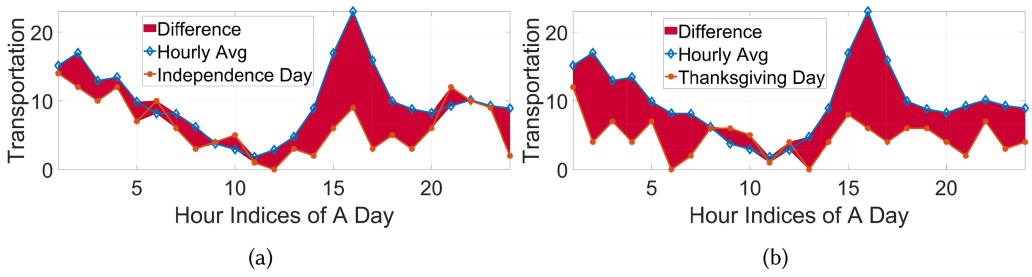


Fig. 7. The citywide crowd activity of transportation category in NYC on *(a)* Independence Day (07/04/2012, Wednesday) and *(b)* Thanksgiving Day (11/22/2012, Thursday).

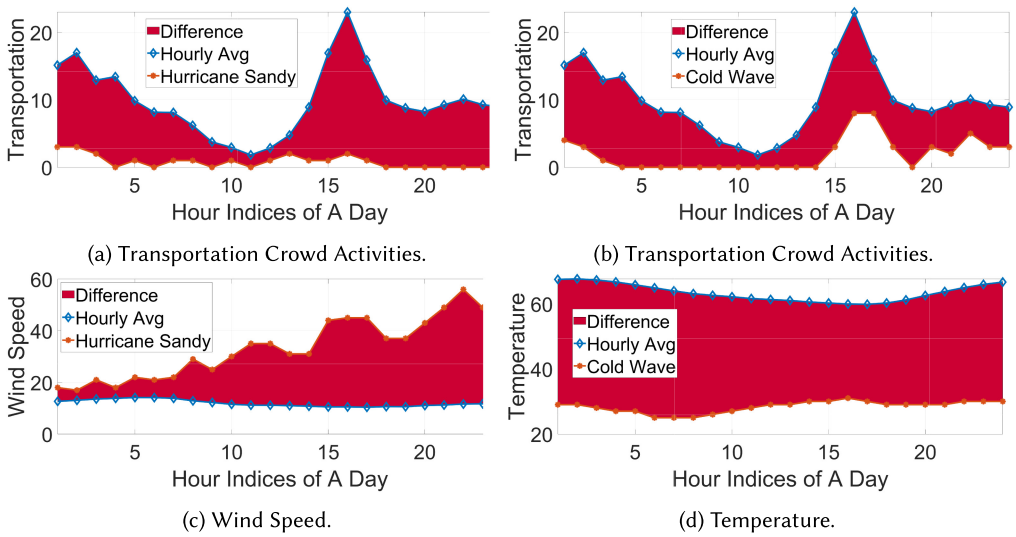


Fig. 8. The citywide crowd activity of transportation category on *(a)* the arrival day of Hurricane Sandy (10/29/2012, Monday); and *(b)* the crowd activities of transportation category. We also show *(c)* the hourly wind speed of NYC on the day of Hurricane Sandy; and *(d)* the hourly temperature of NYC on the arrival day of code wave (02/04/2013, Monday).

and the ones with the impacts of Hurricane Sandy[1] (10/29/2012, Monday) and cold wave (orange) (02/04/2013, Monday) in NYC. We also show in Figures 8(c) and 8(d) the resulting wind speeds and temperature that differed from the historical average. From the differences (highlighted in red), we can observe that the citywide crowd activities of transportation category of NYC went beyond the normal temporal patterns on weekdays and declined with the increased wind speeds and the decreased temperature.

4 SPATIO-TEMPORAL INTERACTIVE RESIDUAL ATTENTION FRAMEWORK FOR CAP

We first overview the structure of STICAP in Section 4.1 and introduce the spatial attention component of STICAP (processing near history, daily, and weekly spatial feature maps) in Section 4.2,

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

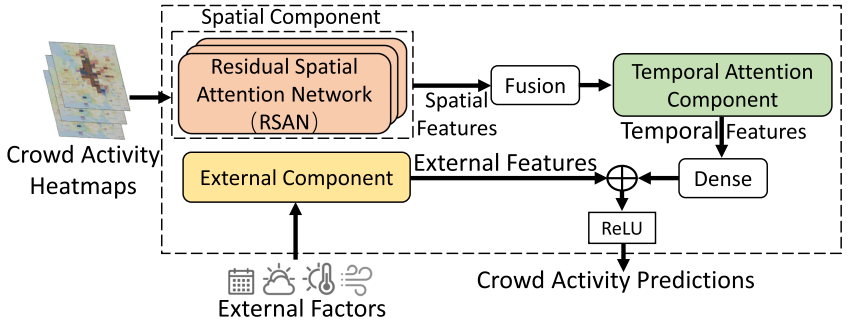


Fig. 9. Illustration of the proposed framework of STICAP.

followed by the structure of temporal attention component in Section 4.3. Finally, we present the structure of external factor component in Section 4.4.

4.1 Model Overview

As presented in Section 3.2, the goal of STICAP is to leverage the historical crowd activity heatmaps and external factors (e.g., weather and holidays) to predict the future occurrences of crowd activities of different categories in one following time step. Figure 9 overviews the architecture of our proposed model STICAP, which contains three major components: *Spatial Attention Component*, *Temporal Attention Component*, and *External Factor Component*.

We first design the *Spatial Attention Component* to capture the spatial distribution and interaction features of the overall crowd activities and the spatial interaction of crowd activities from different regions. In addition, the spatial distributions of the overall crowd activities tend to vary in different time periods. Specifically, the occurrences of the overall crowd activities show explicit hourly, daily, and weekly trends. To leverage this fact, we capture the near history, daily, and weekly spatial features and interactions of the overall crowd activities in the Spatial Attention Component in parallel to gain the general historical spatial distribution patterns and interactions of the overall crowd activities. Within the spatial component, we have designed the *trunk branch,* which functions as the spatial feature processing layer, and the *mask branch*, which works as a spatial feature selection layer to quantify the weights of different regions at different time steps.

In addition, we have designed the *Temporal Attention Component* to capture the temporal features of crowd activities of different categories and differentiate the temporal interactions of crowd activities from different time steps. We also take into account the impacts of external factors on the crowd activities of different categories in the *External Factor Component*.

4.2 Spatial Attention Component

We first utilize the historical heatmaps of all the crowd activities to capture the spatial distribution features of the crowd activities in different time steps. The historical overall crowd activity observations in different regions may have varying impacts on the occurrence of crowd activities of each category at different time periods. To capture such impacts, we propose the ***Residual Spatial Attention Network*** **(RSAN)** to capture the pairwise spatial interactions of the overall crowd activities of all the regions in various historical periods.

In particular, we utilize three parallel RSANs in the Spatial Attention Component to capture the near history, daily, and weekly spatial distribution features of the overall crowd activities in different time steps. After having the above three types of spatial distribution features, we fuse these features by the *Parametric-matrix-based Fusion* to have our final spatial distribution features of the

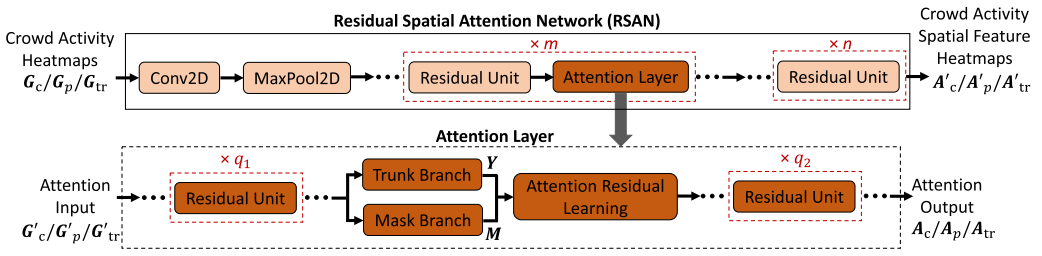


Fig. 10. Residual spatial attention network (RSAN).

overall crowd activities at different time steps. Note that we first apply the Min-Max Normalization to all the crowd activity inputs before further operations. Since the three parallel RSANs in the Spatial Component share the same structures, we will take the learning of the near history spatial distribution features as the example to introduce the design details of the RSAN as follows:

Figure 10 summarizes the main operations inside the RSAN. As shown in Figure 10, given d time steps of near history overall crowd activity heatmaps $\mathbf{G}_c = \{\mathbf{G}_{c,t-L}, \ldots, \mathbf{G}_{c,t-1}\} \in \mathbb{R}^{L \times I \times J}$ during the time period $\{t - L, \ldots, t - 1\}$, the RSAN first captures the preliminary spatial features of the overall crowd activities by a convolution operation *Conv2D*. However, due to the skewness of the crowd activities, crowd activities only co-occur frequently in some of the city regions. To focus on the important spatial features rather that the overall crowd activities, we utilize a MaxPool2D layer to further extract the preliminary spatial distribution feature heatmaps $\mathbf{G'}_c = \{\mathbf{G'}_{c,t-L}, \ldots, \mathbf{G'}_{c,L-1}\} \in \mathbb{R}^{L \times I' \times J'}$, where I' and J' are the height and width of the extracted feature heatmaps. Then, we process the extracted features with m consecutive combinations of *Residual Units* and *Attention Modules* to further extract the deep spatial distribution features and the spatial interaction of crowd activities from different regions.

Finally, we feed the extracted features into n consecutive Residual Units to have the final spatial distributions and features of the RSAN. We overview each components of the RSAN as follows:

(1) *Residual Unit*: The Residual Unit in the RSAN consists of three consecutive identical blocks to capture the preliminary spatial features and interactions of crowd activities. Each identical block includes the **batch normalization (*BN*)**, the *ReLU* activation, and the **convolution operation (*Conv2D*)** as follows:

$$\mathbf{G'}_{c,l} = \mathsf{Conv2D}(\mathsf{ReLU}(\mathsf{BN}(\mathbf{G'}_{c,l}))), \tag{3}$$

where $\mathbf{G'}_{c,l}$ represents the spatial feature heatmap of the overall crowd activities in time step $l \in \{t - L, \ldots, t - 1\}$.

(2) *Attention Layer*: The Attention Layers (as illustrated in Figure 10) in the RSAN are first going through q_1 consecutive Residual Units, followed by two branches, i.e., *Trunk Branch* and *Mask Branch*. The *trunk branch* functions as the spatial feature processing layer and the *mask branch* works as a spatial feature selection layer. We note that the conventional deep convolutional network of attention layer may result in value degradation [32]. To mitigate the degradation problem, the outputs of the Trunk Branch and Mask Branch are fused by an *Attention Residual Learning* module. We utilize the Attention Residual Learning design to weigh the spatial interactions of the overall crowd activities from different regions in each time step, followed by q_2 consecutive Residual Units to achieve the final spatial feature extraction of the Attention Layer.

Details of the two branches in the attention layer and the *Attention Residual Learning* module are further described below.

—**Trunk Branch.** Taking the preliminary spatial distribution feature heatmap of the overall crowd activities, $\mathbf{G'}_c = \{\mathbf{G'}_{c,t-L}, \ldots, \mathbf{G'}_{c,t-1}\} \in \mathbb{R}^{L\times I'\times J'}$, of time steps $\{t-L, \ldots, t-1\}$ as input, the *Trunk Branch* further extracts the deep spatial distribution features $\mathbf{Y} = \{\mathbf{Y}_{t-L}, \ldots, \mathbf{Y}_{t-1}\} \in \mathbb{R}^{L\times I'\times J'}$ of the overall crowd activities. The structure of the Trunk Branch is a sequence of p consecutive Residual Units. The operation of the Trunk Branch on the preliminary spatial distribution heatmap $\mathbf{G'}_{c,l} \in \mathbb{R}^{I'\times J'}$ in time step $l \in \{t-L, \ldots, t-1\}$ is denoted as $\mathbf{Y}_l = \texttt{TrunkBranch}(\mathbf{G'}_{c,l})$.

—**Mask Branch.** With the preliminary spatial distribution heatmap of the overall crowd activities $\mathbf{G'}_c = \{\mathbf{G'}_{c,t-L}, \ldots, \mathbf{G'}_{c,t-1}\} \in \mathbb{R}^{L\times I'\times J'}$ of time steps $\{t-L, \ldots, t-1\}$, the Mask Branch further generates masks $\mathbf{M} = \{\mathbf{M}_{t-L}, \ldots, \mathbf{M}_{t-1}\} \in \mathbb{R}^{L\times I'\times J'}$ to represent and quantify the weights of the spatial distribution features of the overall crowd activities from different regions in each of the time steps during$\{t-L, \ldots, t-1\}$. Specifically, the structure of the Mask Branch is composed of the following two important sub-components, i.e., the *Down-Sampling* and the *Up-Sampling* operations:

(1) *Down-Sampling*: Since the crowd activities are concentrated in some of the city regions, the Down-Sampling sub-component is designed to capture the narrow representative spatial distribution features of the overall crowd activities. In the Down-Sampling sub-component, we first use the `MaxPool2D` operation to reduce the spatial impact of the sparse crowd activities areas. Then, we utilize m consecutive operation combinations to further capture the spatial distribution features. In each operation combination, we perform r consecutive times of Residual Units, which are followed by a `MaxPool2D` operation. After the m consecutive operation combinations, we utilize $2 \times r$ Residual Units to achieve the Down-Sampling operation.

(2) *Up-Sampling*: The Up-Sampling sub-component is proposed to restore the shape of the spatial distribution feature heatmap from the Down-Sampling sub-component and weigh the spatial distribution features of the overall crowd activities from different regions. An `UpSampling2D` operation is used to first recover the preliminary spatial distribution features of the crowd activities in each time step. Another m consecutive operation combinations are further utilized to generate the total spatial distribution features of the crowd activity in each time step. Different from the m consecutive operation combinations from the Down-Sampling sub-component, we perform r consecutive times of Residual Units and one `UpSampling2D` in each operation combinations in the Up-Sampling sub-component. Finally, we utilize the `Conv2D` and the `Softmax` activation function to generate the weight scores $\mathbf{M} = \{\mathbf{M}_{t-L}, \ldots, \mathbf{M}_{t-1}\} \in \mathbb{R}^{L\times I'\times J'}$ for the overall crowd activities from each region in time steps $\{t-L, \ldots, t-1\}$. Specifically, the operation of the Mask Branch on the preliminary spatial distribution feature heatmap in the time step l is given by $\mathbf{M}_l = \texttt{MaskBranch}(\mathbf{G'}_{c,l})$.

—**Attention Residual Learning.** The *Attention Residual Learning* after the Trunk Branch and Mask Branch in the Attention Layer of `RSAN` is used to differentiate the spatial interactions of the overall crowd activities from different regions in each time step and mitigate the degradation problem. With the *Attention Residual Learning*, the output $\mathbf{A}_{c,l} \in \mathbb{R}^{I'\times J'}$ of the Attention Layer in time step l is given by

$$\mathbf{A}_{c,l} = \mathbf{Y}_l \times \mathbf{M}_l, \tag{4}$$

where $\mathbf{M}_l^{(i,j)}$ ranges from 0 to 1. We note that the more $\mathbf{M}_l^{(i,j)}$ approaches 1, the more $\mathbf{A}_{c,l}$ approximates the original spatial distribution features of $\mathbf{G'}_{c,l}$.

After the Attention Residual Learning operation, q_2 consecutive Residual Units are applied to $\mathbf{A}_{c,l}$ and `STICAP` generates the final spatial distribution feature heatmap $\mathbf{A'}_{c,l} \in \mathbb{R}^{I'\times J'}$ of the `RSAN`

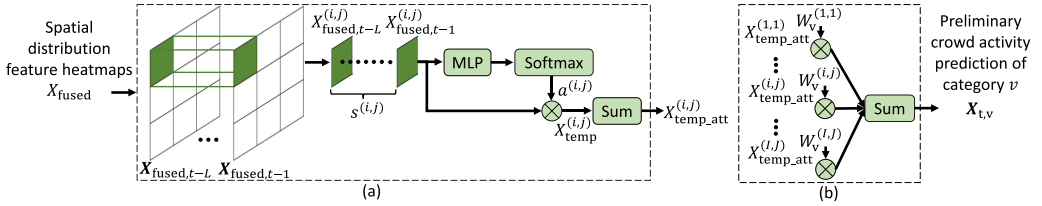


Fig. 11. Designs of temporal components: (a) filter ensemble with the attention weights for the crowd activities in the same region but from different time steps; (b) integration of predictions of crowd activities of category v in time step t.

in time step l. Following the same method, we obtain $\mathbf{A'}_{p,l}$ and $\mathbf{A'}_{tr,l}$ as the daily and weekly spatial feature heatmap, respectively.

—**Parametric-matrix-based Fusion.** After having the near history spatial feature heatmap $\mathbf{A'}_{c,l} \in \mathbb{R}^{I' \times J'}$, daily spatial feature heatmap $\mathbf{A'}_{p,l} \in \mathbb{R}^{I' \times J'}$, and weekly spatial feature heatmap $\mathbf{A'}_{\mathtt{tr},l} \in \mathbb{R}^{I' \times J'}$ of the overall crowd activities in time step l, a *Parametric-matrix-based Fusion* method is used to fuse the three feature heatmaps as the final spatial feature heatmap in time step l by

$$\mathbf{X}_{\text{fused},l} = \mathbf{A'}_{c,l} \circ \mathbf{U}_c + \mathbf{A'}_{p,l} \circ \mathbf{U}_p + \mathbf{A'}_{\mathtt{tr},l} \circ \mathbf{U}_{\mathtt{tr}}, \tag{5}$$

where $\mathbf{X}_{\text{fused},l} \in \mathbb{R}^{I' \times J'}$, $\circ$ is Hadamard product, and $\mathbf{U}_c$, $\mathbf{U}_p$, and $\mathbf{U}_{\mathtt{tr}} \in \mathbb{R}^{I' \times J'}$ are, respectively, the learnable parameters that measure the interactions of the spatial features of the overall crowd activities from the near history, daily, and weekly spatial feature heatmaps of the same region. The fused feature heatmaps in time steps $\{t-L, \ldots, t-1\}$ are denoted as

$$\mathbf{X}_{\text{fused}} = \{\mathbf{X}_{\text{fused},t-L}, \ldots, \mathbf{X}_{\text{fused},t-1}\} \in \mathbb{R}^{L \times I' \times J'}. \tag{6}$$

4.3 Temporal Attention Component

The occurrences of crowd activities show multiple temporal patterns, as illustrated in Figure 6(c). We can further infer that crowd activities under the close temporal context tend to have similar occurrence frequencies. To further extract the temporal interactions among crowd activities, the Temporal Attention Component is proposed to differentiate the impacts of the historical crowd activities from different time steps on the crowd activity predictions of V categories. Figure 11 illustrates the structure of the Temporal Component in predicting the occurrence of crowd activity of category v in time step t.

There are two operations in the Temporal Attention Component, which are *Filter Ensemble* and *Scale Reweight*. Having the spatial feature heatmaps of the overall crowd activities $\mathbf{X}_{\text{fused}} = \{\mathbf{X}_{\text{fused},t-L}, \ldots, \mathbf{X}_{\text{fused},t-1}\} \in \mathbb{R}^{L \times I' \times J'}$ during time steps $\{t-L, \ldots, t-1\}$ from the Spatial Attention Component, the Temporal Attention Component treats the spatial distribution feature heatmaps in different time steps as different temporal layers and weights the importance of each layer.

Let $\mathbf{X}_{\text{fused},l}$ be the spatial feature heatmap in temporal layer $l \in \{t-L, \ldots, t-1\}$. The details of the Filter Ensemble and Scale Reweight are presented as follows:

—**Filter Ensemble.** The overall crowd activities of a specific region in different time steps are viewed as the *descriptors*, where the overall crowd activities $\mathbf{X}^{(i,j)}_{\text{fused},l}$ in region (i,j) denotes the *descriptor* of this region in temporal layer l, where $i \in [1, \ldots, I']$ and $j \in [1, \ldots, J']$. We denote all the descriptors of the same region in different temporal layers as $s^{(i,j)}$, as shown in Figure 11(a).

$$s^{(i,j)} = \left[\mathbf{X}^{(i,j)}_{\text{fused},t-L}, \ldots, \mathbf{X}^{(i,j)}_{\text{fused},t-1}\right]. \tag{7}$$

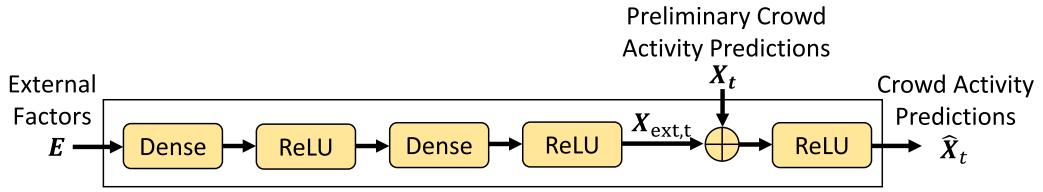


Fig. 12. External factor component.

—**Scale Re-weighting.** We obtain the attention weights $\mathbf{A}^{(i,j)}$ of region (i, j) in all temporal layers by feeding the L descriptors into a **multi-layer perceptron (MLP)** layer followed by a *Softmax* operation as shown as follows:

$$\mathbf{A}^{(i,j)} = \texttt{Softmax}(\texttt{MLP}(s^{(i,j)})), \quad \mathbf{A}^{(i,j)} = \left[\mathbf{A}^{(i,j)}_{t-L}, \ldots, \mathbf{A}^{(i,j)}_{t-1}\right]. \tag{8}$$

The attention value $\mathbf{X}^{(i,j)}_{\texttt{temp},l}$ of region (i, j) at time step $l \in \{t-L, \ldots, t-1\}$ is the product of the attention weight $\mathbf{A}^{(i,j)}_l$ of region (i, j) at temporal layer l and the corresponding indicator $s^{(i,j)}_l$ as:

$$\mathbf{X}^{(i,j)}_{\texttt{temp},l} = s^{(i,j)}_l \times \mathbf{A}^{(i,j)}_l. \tag{9}$$

Given above, the final attention value of region (i, j) is generated by

$$\mathbf{X}^{(i,j)}_{\texttt{temp\_att}} = \sum_{t'=t-L}^{t-1} \mathbf{X}^{(i,j)}_{\texttt{temp},t'}. \tag{10}$$

To generate the non-regional preliminary prediction of V categories of crowd activities in time step t, a linear operation is applied in $\mathbf{X}_{\texttt{temp\_att}} \in \mathbb{R}^{I' \times J'}$ as shown in Figure 11(b) and we have

$$\mathbf{X}_{t,v} = \sum_{i=1,j=1}^{I',J'} \left(\mathbf{W}^{(i,j)}_v \mathbf{X}^{(i,j)}_{\texttt{temp\_att}}\right), \tag{11}$$

where $W^{(i,j)}_v \in \mathbb{R}$ is the weight measuring the importance of the overall crowd activities $\mathbf{X}^{(i,j)}_{\texttt{temp\_att}}$ of region (i, j) on the crowd activity prediction of category $v \in \{1, \ldots, V\}$ in time step t. Thus, the preliminary crowd activity predictions of V categories $\mathbf{X}_t \in \mathbb{R}^V$ in time step t are formed as

$$\mathbf{X}_t = \{\mathbf{X}_{t,1}, \ldots, \mathbf{X}_{t,V}\}. \tag{12}$$

4.4 External Factor Component

The occurrence of crowd activity can also be affected by multiple related external factors, such as weather, weekday, holiday, events, and so on. Therefore, in this study, we design the External Factor Component to further account for the external influence of weather data of temperature, wind speed, and weekdays and holidays on the occurrences of crowd activities. The design of the External Factor Component is illustrated in Figure 12.

Since the weather context is a text description of the observed condition of the weather, we utilize the one-hot encoding method to transfer the weather context into feature vectors. In addition, the temperature and wind speed of the weather data are min-max normalized before being fed into the external factor component. We leverage two separate vectors to represent the weekday/weekend and holiday information, respectively. In the weekday/weekend vector, the vector values are set to be 1 to represent the time step on weekday and 0 for weekend. Similarly in the holiday vector, we set the vector value to be 1 if the time step is on holidays and 0 otherwise. Let

$\mathbf{E}_l \in \mathbb{R}^p$ be the external factor vector in time step $l \in \{t-L, \ldots, t-1\}$, and p is the number of external factors. $\mathbf{E}_l$ contains the min-max normalized wind speed, temperature, the one-hot encoding vector of weather context, and the indicators of weekdays/weekends and holidays at time step l. The external factors during time steps $\{t-L, \ldots, t-1\}$ are formed as $\mathbf{E} \in \mathbb{R}^{l \times p}$.

Our External Factor Component is formed by a two-layer fully connected neural network. It is used for weighting the predicted crowd activities of all V categories from the Temporal Attention Component. It takes $\mathbf{E}_l$ as input, followed by a dense layer for embedding sub-factors, a ReLU activation function, a second dense layer for mapping the low dimension data to high dimension data, and a second ReLU activation function, i.e.,

$$\mathbf{X}_{\text{ext},t} = \text{ReLU}((\text{ReLU}(\mathbf{E} \cdot \mathbf{W}_1 + B_1)) \cdot \mathbf{W}_2 + B_2), \tag{13}$$

where $\mathbf{X}_{\text{ext},t} \in \mathbb{R}^V$ is the crowd activities predictions in time step t from the external factors, $\mathbf{W}_1$, $\mathbf{W}_2$, B_1, and B_2 are weight matrices and bias vectors in the two dense layers, and ReLU represents the activation function, respectively.

With the crowd activity prediction $\mathbf{X}_t$ from the Temporal Attention Component and $\mathbf{X}_{\text{ext},t}$ from the External Factor Component, we obtain the final crowd activity predictions $\mathbf{X} \in \mathbb{R}^V$ in time step t by feeding the merged result (denoted as operator $\bigoplus$) of $\mathbf{X}_{\text{ext},t}$ and $\mathbf{X}_t$ into a ReLU activation operator, i.e.,

$$\hat{\mathbf{X}_t} = \text{ReLU}\left(\mathbf{X}_{\text{ext},t} \bigoplus \mathbf{X}_t\right). \tag{14}$$

5 EXPERIMENTAL STUDIES

We first present the baselines and experimental settings in Section 5.1, followed by the experimental results in Section 5.2.

5.1 Baselines & Experimental Settings

—**Baseline Approaches.** In this study, we compare our proposed method STICAP with the following nine baselines and state-of-art algorithms:

(1) CNN: **Convolutional Neural Network (CNN)** extracts the grid-based check-in features in the input heatmap for the prediction.
(2) RNN/GRU/LSTM: We flatten the grid-based check-in data in each time unit into a 1D vector before feeding them, respectively, into **Recurrent Neural Networks (RNN)**, **Gated Recurrent Unit (GRU)**, and **Long Short-Term Memory (LSTM)** neural network.
(3) CNN+RNN: This baseline combines the CNN with RNN. It leverages the historical grid-based data to predict the future category-based data.
(4) CNN+GRU: CNN is followed by GRU to predict the category-based data.
(5) CNN+LSTM: CNN and LSTM are combined to predict the category-based data.
(6) ConvLSTM: The historical grid-based check-in data are fed into the ConvLSTM [27], followed by a *Flatten* layer and *Dense* layer to generate final prediction.
(7) **Temporal Attention Component (TempAtten)**: We replace the RSAN in the Spatial Attention Component with three identical residual units in Reference [38]. The fused output X_{fused} from the Spatial Attention Component is followed by the Temporal Attention Component.
(8) **Residual Unit & Temporal Attention Component (ResUnit+TempAtten)**: We replace the RSAN with a Residual Unit in Section 4.2. The Temporal Attention Component is added after the fused output X_{fused} from the above modified Spatial Component.
(9) DeepST: The proposed model of DeepST in Reference [38] is adapted as a baseline in this study. In particular, the parameters l_c, l_p, and l_q in DeepST are set as $l_c \in \{3, 4, 5\}$, $l_p \in \{1, 2, 3\}$, and $l_q \in \{1, 2, 3\}$, respectively.

(10) CHAT: which leverages the **Cross-Interaction Hierarchical Attention (CHAT)** network [15] for CAP. The input of CHAT are the near history crowd activity heatmaps $\mathbf{G}_c \in \mathbb{R}^{L\times I\times J}$ during the time period $\{t-L,\ldots,t-1\}$.
(11) ST-Norm: which implements the **Spatial and Temporal Normalization-based framework (ST-Norm)** [10] for CAP.
(12) FDW: which leverages the multi-variate time series forecasting algorithm **Forecast Distance Weighting (FDW)** [6] to take in the historical citywide crowd activities and predict the crowd activities in time step t.

—**Experimental Settings.** In this study, we evaluate our proposed method and the baselines using the Foursquare check-in data of NYC and Tokyo, and the Gowalla check-in data of LA. The Foursquare check-in data are both during the period of 04/02/2012–08/22/2012. The Gowalla check-in data are during the period of 12/01/2009–04/30/2010. The data of last 20 days of each dataset are utilized for testing, the previous 10 days for validation and the others are used for training. We predict 9 categories of crowd activities for NYC and Tokyo and 10 categories for LA.

For the crowd activity heatmaps, all the crowd activities of NYC, Tokyo, and LA are partitioned into 32×16 grid maps. The city maps of NYC, Tokyo, and LA are all approximately squares. We select these grid shapes so the partition granularity of latitude is finer than that of longitude. This is because the activities are concentrated longitudinally, as shown in Figure 4(a). To compare the model performance across different cities, we partition the city maps of the three cities into the same grid shapes. The length of one time step of NYC and Tokyo are set as one hour, while due to the reasons for data sparseness of LA, the length of time unit of LA is set as three hours. L, m, n, q_1, q_2, I, J, I', J', and p are set as 3, 3, 4, 1, 1, 32, 16, 16, 8, and 5, respectively. The learning rate is set as 0.0003, batch size is 128, the data are trained with 500 iterations.

Mean Absolute Error (MAE), Mean Squared Error (MSE), and **Error Rate (ER)** are utilized to evaluate the performance of the algorithms, i.e.,

$$\text{MAE}=\frac{1}{V}\times\sum_{v=1}^{V}|\hat{\mathbf{X}}_{v,t}-\mathbf{X}_{v,t}|,\quad \text{MSE}=\frac{1}{V}\times\sum_{v=1}^{V}\left(\hat{\mathbf{X}}_{v,t}-\mathbf{X}_{v,t}\right)^2,\quad \text{and}\quad \text{ER}=\frac{\sum_{v=1}^{V}|\hat{\mathbf{X}}_{v,t}-\mathbf{X}_{v,t}|}{\sum_{v=1}^{V}\mathbf{X}_{v,t}},\tag{15}$$

where $\hat{\mathbf{X}}_{v,t}$ denotes the predicted value of crowd activity category v in time step t. $\mathbf{X}_{v,t}$ denotes the ground truth value of a crowd activity category v in time step t. The models are trained based on the loss of MSE. All the experiments are conducted upon a desktop server with Intel i7-9700, NVIDIA GeForce RTX 2060 SUPER, 16.0 GB RAM, and Windows 10.

5.2 Evaluation Results

—**General Performance.** As demonstrated in Table 3, our proposed STICAP has achieved the overall better performance than the other baselines, with an average error reduction of 35.02%. The results show the effectiveness of STICAP designs on handling the non-regional crowd activity prediction problem with check-in data.

We briefly review and compare other baseline approaches. CNN extracts the shallow spatial distribution features of crowd activities, and ConvLSTM captures the preliminary temporal patterns. Therefore, neither CNN nor ConvLSTM performs well in the CAP. CHAT considers the general pairwise spatial, temporal, and categorical relationship among crowd activities for CAP. However, we note that crowd activities are often impacted by the external factors (e.g., the extreme weathers and special holidays) and their activity levels may fluctuate beyond the general patterns. RNN, GRU, LSTM, and TempAtten focus only on the temporal patterns and therefore cannot fully capture the spatial features within the crowd activities. Despite the designs of learning both the spatial and

Table 3. Prediction Results and Performance Comparison on NYC, Tokyo, and LA

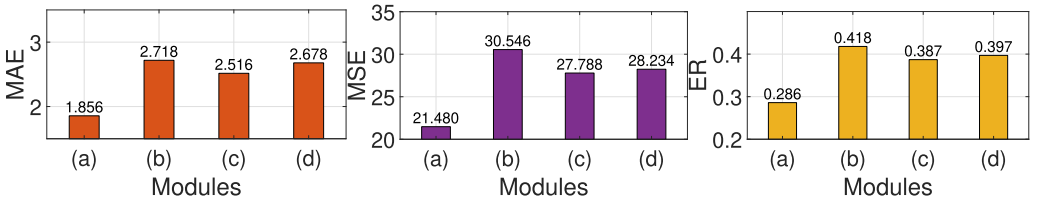
Scheme	NYC (9)			Tokyo (9)			LA (10)		
	MAE	MSE	ER	MAE	MSE	ER	MAE	MSE	ER
CNN	1.691	8.851	0.725	4.276	116.765	0.694	0.944	3.673	0.773
RNN	1.636	11.157	0.701	3.822	77.015	0.620	1.116	4.922	0.914
GRU	1.769	11.256	0.758	3.153	78.260	0.512	0.995	3.761	0.815
LSTM	1.542	8.910	0.661	3.636	77.507	0.773	0.977	4.08	0.800
ConvLSTM	1.549	8.122	0.630	3.838	120.751	0.591	1.250	5.626	0.909
CNN+RNN	1.757	9.838	0.753	3.185	74.480	0.517	0.971	3.382	0.789
CNN+GRU	1.742	8.871	0.746	2.998	60.113	0.672	0.955	3.436	0.776
CNN+LSTM	1.727	9.523	0.740	2.899	60.326	0.470	0.981	3.632	0.797
TempAtten	1.307	5.337	0.601	3.139	29.980	0.483	0.858	2.879	0.624
ResUnit+TempAtten	1.447	5.155	0.589	2.288	28.340	0.352	0.847	2.952	0.615
DeepST	1.662	8.002	0.676	3.902	67.440	0.601	1.110	5.116	0.807
CHAT	1.679	10.118	0.683	3.406	107.766	0.593	1.174	5.404	0.854
ST-Norm	1.533	7.464	0.624	3.363	63.458	0.518	0.903	3.152	0.682
FDW	1.354	5.633	0.545	2.342	29.034	0.373	0.923	3.532	0.702
STICAP	1.192	5.460	0.484	1.856	21.480	0.286	0.865	2.706	0.629

temporal crowd mobility, we observe that DeepST may not necessarily capture the large-scale spatial features with different spatial focuses and with different granularity levels. With the spatial and temporal attention mechanisms, our proposed STICAP can capture the most relevant interactions from spatial and temporal dimensions based on the context of the prediction, thus outperforming other approaches. ST-Norm considers factorizing the crowd activities but does not account for the geospatial distributions among regions. FDW focuses on the small subset of variables available within the time series and does not account for learning the spatial interactions among the time-series variables and hence may not adapt to the complex crowd activities. Different from these approaches, our STICAP not only accounts for multiple temporal patterns among crowd activities, but also captures the spatial interactive distribution of crowd activities, thus yielding the better accuracy in CAP.

Comparing the prediction results on the NYC and Tokyo datasets in Table 3, we can also see that prediction errors of Tokyo are relatively higher than those of NYC and LA. It is likely because the check-in data of NYC and LA are much sparser than the data of Tokyo. Specifically, the overall hourly frequencies of crowd activity of Tokyo are much greater than those of NYC and LA. The range of hourly crowd activity of each venue category of Tokyo is from 0 to 194, while that of NYC is between 0 and 38, and that of LA is between 0 and 26. While the settings of TempAtten and ResUnit+TempAtten marginally outperform STICAP for the LA dataset (likely due to easier learning on the sparse data for simple attention designs), STICAP still substantially outperforms many other state-of-the-art approaches.

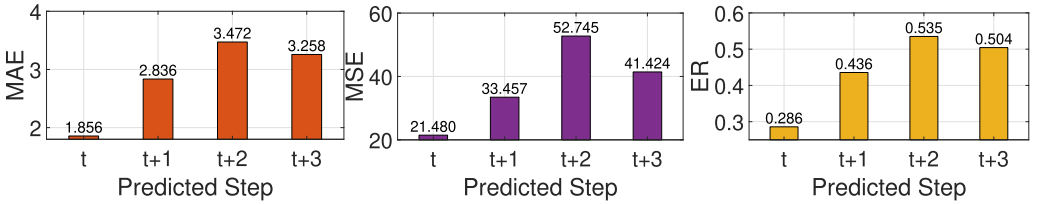
—**Ablation Studies.** We have conducted ablation studies to evaluate the importance for the occurrence of crowd activities by evaluating each component of STICAP using the data of Tokyo as follows (by comparing the complete STICAP that is labeled as (a)):

(b) *Spatial Attention Component*: To apply the spatial Attention Component as introduced in Section 4.2 alone for crowd activities prediction, the spatial attention output of $\mathbf{X}_{\text{fused}}$ in Equation (5) is modified as category format by being fed into a dense layer with the output sequence of V.



(a) MAE. (b) MSE. (c) ER.

Fig. 13. Model ablation results of Tokyo predicting nine categories of crowd activities using (a) STICAP; (b) spatial attention module; (c) spatial attention module+external factor module; and (d) replacing the temporal attention module in STICAP as multiple GRU layers.



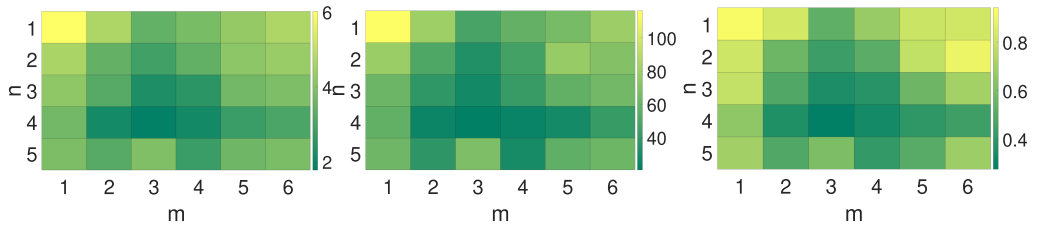
(a) MAE. (b) MSE. (c) ER.

Fig. 14. Multi-step prediction results of Tokyo predicting the crowd activities of V categories in time step t, $t+1$, $t+2$, and $t+3$, respectively.

(c) *Spatial Attention Component & External Factor Component*: The Spatial Attention Component and the External Factor Component are combined to predict crowd activities of each category.
(d) GRU *Instead of Temporal Attention Module*: We replaced the Temporal Attention Module in STICAP by multiple GRU layers.

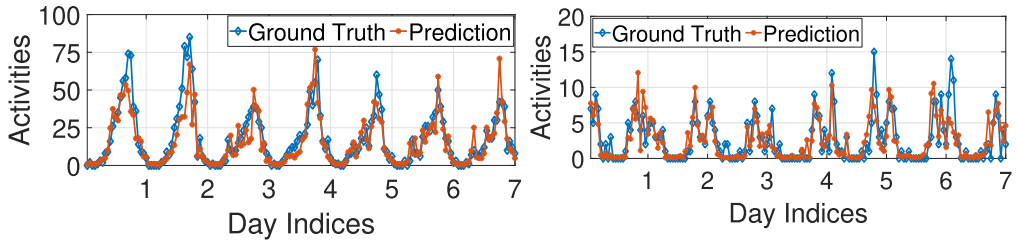
As shown in Figure 13, by using all the components of STICAP, the MAE and ER prediction results greatly improve in the prediction of nine categories of crowd activities. In STICAP, the Spatial Attention Module captures the hourly, daily, and weekly deep spatio-temporal distribution features of crowd activities. Therefore, the Spatial Attention Module works well alone in CAP. By further accounting for the impacts from the external factors on crowd activities, STICAP improves the prediction accuracy when the crowd activities fluctuate. GRU performs well in capturing the short-term temporal correlations. However, it may not fully capture the data with long-term dependencies. The prediction result demonstrates the effectiveness of STICAP in urban CAP.

—**Sensitivity Analysis.** To further verify our proposed method STICAP on multi-step prediction, we predict the crowd activities of V categories in time step t, $t+1$, $t+2$, and $t+3$, respectively. The prediction results are shown in Figure 14. We can see that the error increases as the STICAP is predicting the longer horizon of the crowd activities. However, STICAP achieves overall robust performance in the CAP. We further examine the influence of parameter setting in RSAN on prediction results. We note that in Figure 10 the Attention Layer in RSAN consists of m consecutive pairs of combined Residual Units and Attention Layers and n consecutive Residual Units after the combinations. To analyze the influence of the values of m and n on the CAP, we have conducted the experiments based on the Tokyo dataset to verify different value combinations of m (ranges from 1 to 6) and n (ranges from 1 to 5). Figure 15 shows the prediction results of the experiments. We can observe that the darker colors imply the smaller prediction errors. We can learn from the figures



(a) MAE. (b) MSE. (c) ER.

Fig. 15. Sensitivity studies on the m (numbers of residual units and attention layers) and n (numbers of consecutive Residual Units).



(a) Shop & service. (b) Food.

Fig. 16. Hourly crowd activity prediction results and ground truth of (a) shop & service of Tokyo from 08/03/2012–08/09/2012, and (b) food of LA from 04/11/2010–04/17/2010.

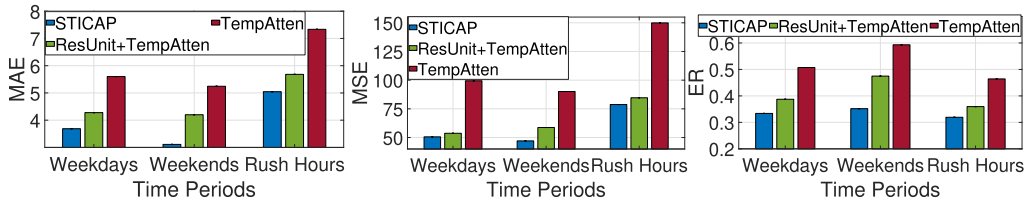
that STICAP achieves the overall better performance with the combinations around $(m, n) = (4, 3)$. Therefore, we set $(m, n) = (4, 3)$ by default.

RSAN is designed to capture the deep spatial distribution features of the crowd activities. Different layers of the Residual Unit and Attention Layer can capture the deep spatial distribution features with focuses on different locations and with various levels of details. However, the over-complicated model and the over-sampled spatial features will as a return decrease the model effectiveness due to the problem of over-fitting.

—**Visualization.** Figure 16 shows the predictions and ground-truths of crowd activities of shop & service of Tokyo and crowd activities of food of LA running with STICAP. The predictions and ground-truths of Tokyo are from 08/03/2012–08/09/2012, and the data from LA are from 04/11/2010–04/17/2010. We can learn from the figures that the overall gaps between predictions and the corresponding ground truths are very small, which demonstrates the high accuracy of STICAP in the prediction.

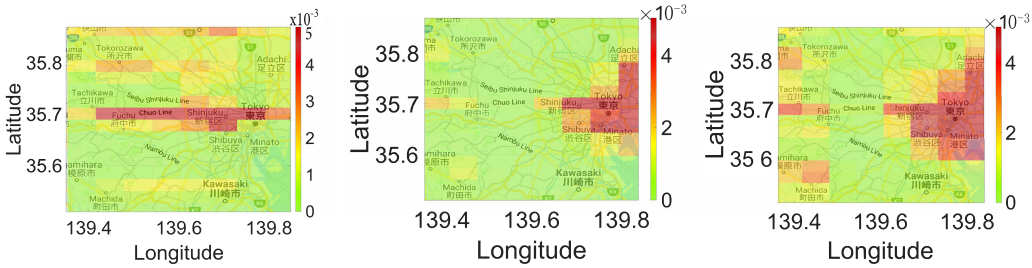
Figure 17 further shows the prediction results during weekdays, weekends, and rush hours in Tokyo by comparing ResUnit+TempAtten, TempAtten with STICAP. As shown in Figure 17, STICAP performs the best in all cases of weekdays, weekends, and rush hours crowd activity predictions. These figures demonstrate the high accuracy of STICAP and its significant improvement from the state-of-the-arts.

To illustrate the effectiveness of the spatial feature extraction and spatial interaction characterization of RSAN, we visualize three scoring matrices generated from the Mask Branches of three RSANs of the Spatial Attention Component when predicting one selected time step of Tokyo testing data, as shown in Figure 18. The scoring matrices heatmaps represent the spatial features of the overall crowd activities of the hourly, daily, and weekly trends. Note that the sum of one scoring



(a) MAE. (b) MSE. (c) ER.

Fig. 17. The MAE, MSE, and ER in weekdays, weekends, and holidays of Tokyo, with STICAP, ResUnit+TempAtten, and TempAtten.



(a) Hourly. (b) Daily. (c) Weekly.

Fig. 18. The selected scoring matrices generated from the Mask Branches of three RSANs of the Spatial Attention Component when predicting one selected time step of Tokyo testing data. The scoring matrices heatmaps represent the spatial features and the interaction levels of the overall crowd activities of the hourly, daily, and weekly trends.

matrix is 1. We can see from the figures that the RSAN can capture the important regions of the overall crowd activities with different temporal trends.

6 DISCUSSION

We briefly discuss the deployment of STICAP in the following aspects:

–*Potential Extensions to Other Applications*: In this work, we aim at crowd activity participation prediction by leveraging the city-wide spatial and temporal characteristics of the location-based social network check-in data and external factors such as weather and weekends/holidays. Based on the prediction of crowd activity participation in a specific time step, our proposed STICAP can be further utilized to detect the citywide abnormal user activities and the corresponding locations.

–*Location Privacy Discussion*: Location-and-timestamp-tagged user activity data are very useful sources to understand the movement and the activity status of crowds. The availability of the crowd activity data enables many research directions, including crowd mobility prediction, crowd flow prediction, and so on. However, privacy issue arises for crowds with the utilization of crowd activity data. We note all user IDs in this study have been hashed into global identifiers by the Foursquare data provider and Gowalla data provider.

7 CONCLUSION

We propose STICAP with spatio-temporal attention mechanism to predict the occurrences of city-wide crowd activities. We have conducted extensive and comprehensive analysis on the check-in data of New York City, Tokyo, and LA. The data analysis also shows weather, weekdays, and

holidays have multi-level impacts on the occurrence of crowd activities. We take into account the impacts of near history, daily, and weekly occurrences of crowd activities on future occurrences of crowd activities. We have designed the Residual Spatial Attention Module RSAN in the Spatial Component to capture and differentiate the spatial interactions of the crowd activities. The Temporal Attention Component is also proposed to differentiate the contributions of the historical crowd activities from different time steps and their interactions with the prediction results. The results have demonstrated that STICAP outperforms other baselines with 35.02% improvement, on average, for NYC, Tokyo, and LA.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] 2009–2010. Gowalla Data. Retrieved from https://go.gowalla.com/
[2] 2012. Foursquare Data. Retrieved from https://foursquare.com/
[3] 2021. Weather Data API. Retrieved from https://api.weather.com/
[4] Manuele Barraco, Nicola Bicocchi, Marco Mamei, and Franco Zambonelli. 2021. Forecasting parking lots availability: Analysis from a real-world deployment. In *IEEE PerCom Workshops*. IEEE, 299–304.
[5] Hancheng Cao, Jagan Sankaranarayanan, Jie Feng, Yong Li, and Hanan Samet. 2019. Understanding metropolitan crowd mobility via mobile cellular accessing data. *ACM Trans. Spatial Algor. Syst.* 5, 2 (2019), 1–18.
[6] Jatin Chauhan, Aravindan Raghuveer, Jay Nandy, Rishi Saket, and Balaraman Ravindran. 2022. Multi-variate time series forecasting on variable subsets. In *ACM SIGKDD*.
[7] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *IEEE CVPR*. 5659–5667.
[8] Gabriele Civitarese, Sergio Mascetti, Alberto Butifar, and Claudio Bettini. 2019. Automatic detection of urban features from wheelchair users' movements. In *IEEE PerCom*. IEEE, 1–10.
[9] Henriette Cramer, Mattias Rost, and Lars Erik Holmquist. 2011. Performing a check-in: Emerging practices, norms and "conflicts" in location-sharing using Foursquare. In *ACM MobileHCI*. 57–66.
[10] Jinliang Deng, Xiusi Chen, Renhe Jiang, Xuan Song, and Ivor W. Tsang. 2021. ST-Norm: Spatial and temporal normalization for multi-variate time series forecasting. In *ACM SIGKDD*. 269–278.
[11] Zipei Fan, Xuan Song, Ryosuke Shibasaki, and Ryutaro Adachi. 2015. CityMomentum: An online approach for crowd behavior prediction at a citywide level. In *ACM UbiComp*. 559–569.
[12] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2018. Soft+ hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection. *Neural Netw.* 108 (2018), 466–478.
[13] Samuli Hemminki, Keisuke Kuribayashi, Shin'ichi Konomi, Petteri Nurmi, and Sasu Tarkoma. 2019. Crowd replication: Sensing-assisted quantification of human behavior in public spaces. *ACM Trans. Spatial Algor. Syst.* 5, 3 (2019), 1–34.
[14] Minh X. Hoang, Yu Zheng, and Ambuj K. Singh. 2016. FCCF: Forecasting citywide crowd flows based on big data. In *ACM SIGSPATIAL*. 1–10.
[15] Chao Huang, Chuxu Zhang, Peng Dai, and Liefeng Bo. 2021. Cross-interaction hierarchical attention networks for urban anomaly prediction. In *IJCAI*. 4359–4365.
[16] Huiqun Huang, Xi Yang, and Suining He. 2021. Multi-head spatio-temporal attention mechanism for urban anomaly event prediction. *Proc. ACMInteract., Mob., Wear. Ubiq. Technol.* 5, 3 (2021), 1–21.
[17] Tomoharu Iwata, Hitoshi Shimizu, Futoshi Naya, and Naonori Ueda. 2017. Estimating people flow from spatiotemporal population data via collective graphical mixture models. *ACM Trans. Spatial Algor. Syst.* 3, 1 (2017), 1–18.
[18] Renhe Jiang, Xuan Song, Dou Huang, Xiaoya Song, Tianqi Xia, Zekun Cai, Zhaonan Wang, Kyoung-Sook Kim, and Ryosuke Shibasaki. 2019. Deepurbanevent: A system for predicting citywide crowd dynamics at big events. In *ACM SIGKDD*. 2114–2122.
[19] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction.. In *IJCAI*. 3428–3434.
[20] Lingbo Liu, Ruimao Zhang, Jiefeng Peng, Guanbin Li, Bowen Du, and Liang Lin. 2018. Attentive crowd flow machines. In *ACM MM*. 1553–1561.
[21] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. 2019. ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding. In *IEEE CVPR*. 3225–3234.

[22] Wenjie Pei, Tadas Baltrusaitis, David M. J. Tax, and Louis-Philippe Morency. 2017. Temporal attention-gated model for robust sequence classification. In *IEEE CVPR*. 6730–6739.
[23] Sanjay Purushotham and C.-C. Jay Kuo. 2016. Personalized group recommender systems for location-and event-based social networks. *ACM Trans. Spatial Algor. Syst.* 2, 4 (2016), 1–29.
[24] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971* (2017).
[25] Julian A. Ramos Rojas, Johana Rosas, Yilin Shen, Hongxia Jin, and Anind K. Dey. 2020. Activity recommendation: Optimizing life in the long term. In *IEEE PerCom*. IEEE, 1–10.
[26] Amin Sadri, Flora D. Salim, Yongli Ren, Wei Shao, John C. Krumm, and Cecilia Mascolo. 2018. What will you do for the rest of the day? An approach to continuous trajectory prediction. *Proc. ACMInteract., Mob., Wear. Ubiq. Technol.* 2, 4 (2018), 1–26.
[27] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214* (2015).
[28] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2014. Prediction of human emergency behavior and their mobility following large-scale disaster. In *ACM SIGKDD*. 5–14.
[29] Douglas Do Couto Teixeira, Aline Carneiro Viana, Jussara M. Almeida, and Mrio S. Alvim. 2021. The impact of stationarity, regularity, and context on the predictability of individual human mobility. *ACM Trans. Spatial Algor. Syst.* 7, 4 (2021), 1–24.
[30] Dimitrios Tomaras, Ioannis Boutsis, and Vana Kalogeraki. 2018. Modeling and predicting bike demand in large city situations. In *IEEE PerCom*. IEEE, 1–10.
[31] Rohit Verma, Bivas Mitra, and Sandip Chakraborty. 2019. Avoiding stress driving: Online trip recommendation from driving behavior prediction. In *IEEE PerCom*. IEEE, 1–10.
[32] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. In *IEEE CVPR*. 3156–3164.
[33] Haiquan Wang, Yilin Li, Guoping Liu, Xiang Wen, and Xiaohu Qie. 2019. Accurate detection of road network anomaly by understanding crowd's driving strategies from human mobility. *ACM Trans. Spatial Algor. Syst.* 5, 2 (2019), 1–17.
[34] Senzhang Wang, Hao Miao, Hao Chen, and Zhiqiu Huang. 2020. Multi-task adversarial spatial-temporal networks for crowd flow prediction. In *ACM CIKM*. 1555–1564.
[35] Dongfang Yang, Ümit Özgüner, and Keith Redmill. 2020. A social force based pedestrian motion model considering multi-pedestrian interaction with a vehicle. *ACM Trans. Spatial Algor. Syst.* 6, 2 (2020), 1–27.
[36] Dingqi Yang, Daqing Zhang, Vincent W. Zheng, and Zhiyong Yu. 2014. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Trans. Syst., Man, Cybern.: Syst.* 45, 1 (2014), 129–142.
[37] Zheng Yang, Feng Wang, and Wei Gong. 2021. Mobi-Track: Distilling direct path in time domain for high accuracy WiFi tracking. In *IEEE PerCom Workshops*. IEEE, 287–292.
[38] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*. AAAI Press, 1655–1661.
[39] Jason Shuo Zhang, Mike Gartrell, Richard Han, Qin Lv, and Shivakant Mishra. 2019. GEVR: An event venue recommendation system for groups of mobile users. *Proc. ACMInteract., Mob., Wear. Ubiq. Technol.* 3, 1 (2019), 1–25.
[40] Xian Zhou, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. 2018. Predicting multi-step citywide passenger demands using attention-based neural networks. In *ACM WSDM*. 736–744.

Received 11 April 2022; revised 6 December 2022; accepted 24 April 2023