Multi-Head Spatio-Temporal Attention Mechanism for Urban Anomaly Event Prediction

HUIQUN HUANG, University of Connecticut, USA
XI YANG, University of Connecticut, USA
SUINING HE, University of Connecticut, USA

Timely forecasting the urban anomaly events in advance is of great importance to the city management and planning. However, anomaly event prediction is highly challenging due to the sparseness of data, geographic heterogeneity (*e.g.,* complex spatial correlation, skewed spatial distribution of anomaly events and crowd flows), and the dynamic temporal dependencies.

In this study, we propose M-STAP, a novel **M**ulti-head **S**patio-**T**emporal **A**ttention **P**rediction approach to address the problem of multi-region urban anomaly event prediction. Specifically, M-STAP considers the problem from three main aspects: *(1)* extracting the spatial characteristics of the anomaly events in different regions, and the spatial correlations between anomaly events and crowd flows; *(2)* modeling the impacts of crowd flow dynamic of the most relevant regions in each time step on the anomaly events; and *(3)* employing attention mechanism to analyze the varying impacts of the historical anomaly events on the predicted data. We have conducted extensive experimental studies on the crowd flows and anomaly events data of New York City, Melbourne and Chicago. Our proposed model shows higher accuracy (41.91% improvement on average) in predicting multi-region anomaly events compared with the state-of-the-arts.

CCS Concepts: • **Information systems** → *Mobile information processing systems.*

Additional Key Words and Phrases: anomaly event prediction, crowd flow, multi-head self-attention

ACM Reference Format:
Huiqun Huang, Xi Yang, and Suining He. 2021. Multi-Head Spatio-Temporal Attention Mechanism for Urban Anomaly Event Prediction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 104 (September 2021), 21 pages. https://doi.org/10.1145/3478099

1 INTRODUCTION

Urban anomaly events are the unusual events which may incur or be accompanied by the abnormal movement of crowd flows. For example, the anomaly events like vehicle collision, noise incident, crime, and service request in urban cities are the typical factors leading to or coming with the sudden and long-term change of crowd flows around the location of event. The occurrence of these anomaly events often threatens public property and security, driving the demand for urban anomaly events evaluation and prediction. With the advances in large scale computing infrastructures, a diverse of spatial and temporal data associated with urban crowd flow dynamics and urban anomaly events are generated in large scale. Using these datasets to predict the occurrence of urban anomaly event in advance can significantly improve the city safety management, risk assessment, traffic management and emergency planning and procedures.

Authors' addresses: Huiqun Huang, huiqun.huang@uconn.edu, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, USA; Xi Yang, xi.yang@uconn.edu, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, USA; Suining He, suining.he@uconn.edu, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, USA.

2474-9567/2021/9-ART104 $15.00
https://doi.org/10.1145/3478099

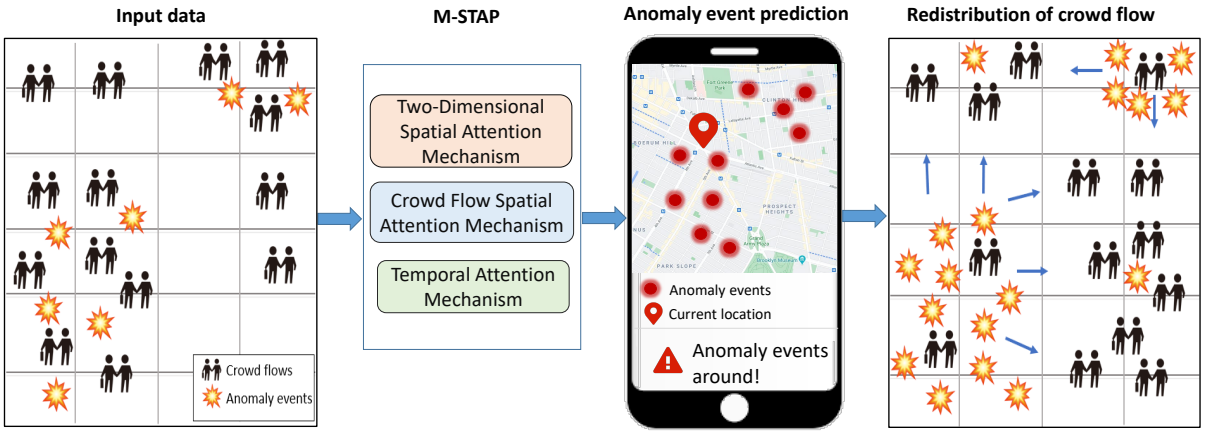


Fig. 1. Motivations of M-STAP in urban anomaly prediction.

Fig. 1 illustrates an example of urban anomaly event prediction application. In the leftmost sub-figure of Fig. 1, the regions with anomaly events are usually more crowded than other regions when in the absence of or in the early stage of anomaly events. By learning and associating the historical data of crowd flows and anomaly events, we aim at *predicting the incoming anomaly event distributions in specific area ahead of time*. Users will receive alerts where the anomaly events will occur, and then be *steered* towards the surrounding safe or shelter areas (the upper left and lower right regions in the leftmost sub-figure of Fig. 1).

To predict the anomaly events and enable the aforementioned crowd redistribution, it is essential to learn the spatio-temporal crowd flow features. Nevertheless, the spatial and temporal *correlations* between various anomaly events, different crowd flows, and anomaly events and crowd flows can be very complex. Along with the occurrence of anomaly events, the crowd flow distribution varies in different regions and different stages of the anomaly events. From the spatial perspective, the crowd flow movements in the surrounding regions of anomaly event locations can always indicate the occurrence of the events. However, the range of the surrounding regions and the degree of impact by the crowd flows on the anomaly events are often hard to model.

On the other hand, the incurred movement of the crowd flows may happen in different stages of the anomaly events. Capturing the spatio-temporal correlations of crowd flows and anomaly events between different regions can be very useful tool for anomaly event prediction. Such correlations between crowd flows and anomaly events also exist in the anomaly events from multiple regions. The co-occurrence of the anomaly events in a single region can be influenced by the anomaly events from other regions with different degrees. At different periods the correlations may vary. Capturing the spatial influence scale and the corresponding influence degree can be a highly challenging task. In addition, predicting multi-region anomaly events is very challenging due to *sparsity* in the crowd flows and anomaly events. The limited observation will significantly degrade the prediction accuracy.

To address the above challenges, we propose M-STAP, a Multi-head Spatio-Temporal Attention Prediction approach for multi-region anomaly events. The proposed model *jointly* takes in the heatmaps of recent historical crowd flows and anomaly events as *inputs*, and *outputs* the future occurrences of anomaly events in each region of a city. In particular, we have designed three main modules within M-STAP, as illustrated in Fig. 1, *i.e.,* **(1)** *Two-dimensional Spatial Attention Mechanism* (TDSAM) to model the spatial characteristics of anomaly events and crowd flows; **(2)** *Crowd Flow Spatial Attention Mechanism* (CFSAM) to select the crowd flows at the most relevant regions in each time step to calibrate the anomaly events of each region; and **(3)** *Temporal Attention Mechanism* (TAM) to extract the temporal patterns of the anomaly events and differentiate the impacts of anomaly events from different historical time steps.

Our main contributions in M-STAP are summarized as follows:

(a) *Learning spatial-temporal correlations between anomaly events & crowd flows*: In this study, we consider the spatial dependencies from two aspects, namely the spatial dependencies among anomaly events from different regions, and the spatial correlations between anomaly events and crowd flows. We also take into account the temporal correlations of the anomaly events from different time steps.

(b) *Predicting multi-region anomaly events by multi-head spatio-temporal attention*: To capture the aforementioned spatio-temporal dependencies, we propose a multi-region anomaly events prediction model, M-STAP. M-STAP extracts the spatial feature maps of anomaly events by concatenating the global spatial feature maps from *Multi-Head Self-Attention Mechanism* (MHSAM) and the local spatial feature maps from traditional Convolution Operation in TDSAM, using both the anomaly event data and crowd flow data. Specifically, each attention head in MHSAM focuses on extracting the spatial correlations of anomaly events among all regions given the citywide anomaly event and crowd flow data.

(c) *Extensive experimental studies on real-world crowd flow data & anomaly events*: To evaluate our proposed model, we conduct extensive experiments on the Foursquare check-in data, 311 noise complaint data, 311 service request data, vehicle collision data and crime complaint data from New York City (NYC), pedestrian counting data and parking data from Melbourne, and bike-sharing usage data and crime data from the City of Chicago. The results show that our proposed model outperforms other baselines in multi-region anomaly event prediction and achieves 41.91% improvement on average compared with the baselines.

• **Technical & Societal Implications**: Our proposed work brings the following societal and technical implications and potential benefits for real-world applications. *(1)* The proposed approach leverages the urban mobility data, which can be easily and timely harvested from the existing mobile devices or social networks, to support proactive and accurate monitoring of the urban anomalies. *(2)* The novel model designs leveraging the novel attention mechanisms can benefit existing urban mobility studies [4, 9, 22, 37], addressing the complex correlations between different urban features and supporting subsequent more proactive decision makings and responses. *(3)* By incorporating our M-STAP into the existing urban mobility platforms [9] such as the mobile map services, car pooling [12], and bike-sharing [13], we can enhance their robustness and adaptivity under urban anomaly events. *(4)* The multi-region anomaly prediction approach and data analytics insights can assist the city planners and urban computing practitioners in designing and implementing related mobile and ubiquitous systems in handling massive crowds and urban anomalies, balancing the resource allocation [25], and advancing the citywide socioeconomic evolution [6].

The rest of the paper is organized as follows. We first overview the datasets in Sec. 2, followed by the problem formulation and the details of our proposed core method M-STAP in Sec. 3. We present our experimental results in Sec. 4. Afterwards, we will review the related work in Sec. 5, and finally conclude in Sec. 6.

2 DATA OVERVIEW FOR M-STAP

In this section, we first overview the dataset used for M-STAP's learning in Sec. 2.1, followed by the data analysis in Sec. 2.2.

2.1 Dataset Overview

In this work, we study on datasets from three cities, New York City (NYC) and City of Chicago, US, and Melbourne, Australia. We summarize the datasets in Tab. 1, and provide an example of one motor vehicle collision record in Tab. 2. We briefly present the datasets as follows.

• *Crowd Flow/Mobility Data*: We have collected the NYC check-in, Melbourne pedestrian counting, and Chicago bike-sharing usage for our model inputs. Specifically, we have:

Table 1. Crowd flow and anomaly event datasets from NYC, Melbourne, and Chicago.

City	Data Type	Data	Description	Geographic Bounding Box Range	Total Records
NYC	Crowd Flow	Check-ins	User id, latitude, longitude, timestamp	[40.55085°N, 40.98833°N], [73.68382°W, 74.27476°W]	115,394
	Anomaly Event	Motor Vehicle Collisions	Latitude, longitude, timestamp, vehicle types, casualties	[40.55085°N, 40.91288°N], [73.70055°W, 74.25453°W]	24,468
		Crime Complaints	Latitude, longitude, timestamp, offense type	[40.55085°N, 40.91504°N], [73.68478°W, 74.25493°W]	204,946
		311 Service Requests	Latitude, longitude, timestamp, requested service type	[40.55094°N, 40.91134°N], [73.70140°W, 74.25208°W]	82,608
		311 Noise Complaints	Latitude, longitude, timestamp	[40.55087°N, 40.91104°N], [73.70223°W, 74.25407°W]	95,217
Melbourne	Crowd Flow	Pedestrian Counts	Latitude, longitude, timestamp, hourly counts	[37.80759°S, 37.82044°S], [144.96142°E, 144.97489°E]	67,108,088
	Anomaly Event	On-street Car Parking	Latitude, longitude, timestamp, street name	[37.80759°S, 37.82044°S], [144.96142°E, 144.97489°E]	9,171,718
Chicago	Crowd Flow	Bike Usage	Latitudes, longitudes, timestamps, station ids of start station and end station of a trip	[41.73664°N, 42.06399°N], [87.54938°W, 87.80287°W]	5,029,240
	Anomaly Event	Crime Events	Latitude, longitude, timestamp, crime type	[41.73664°N, 42.06399°N], [87.54938°W, 87.80287°W]	117,018

Table 2. An example of anomaly event (motor vehicle collision).

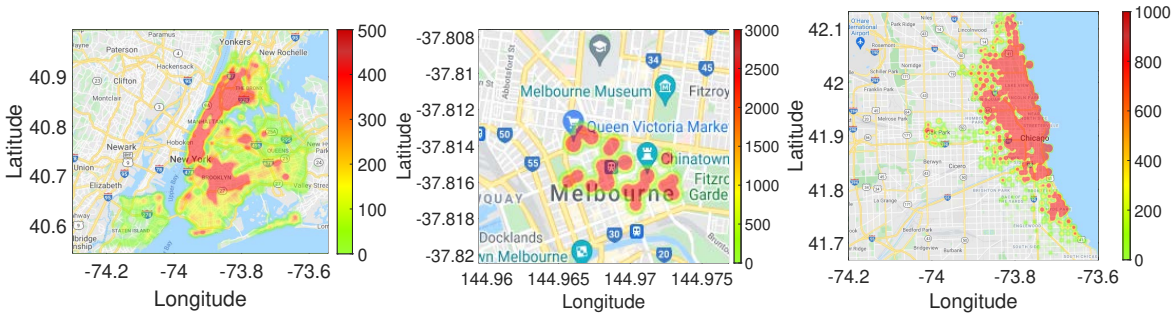
Latitude	Longitude	UTC Timestamp	Vehicle type	Casualties
40.810318°N	73.943634°W	12/14/2018 18:45:00	Station Wagon/Sport Utility Vehicle	0

(1) **NYC Check-in Data:** In this study, we utilize the check-in data of NYC during 2012/04/02 – 2012/08/22 to indicate the movement of crowd flow. There are totally 115,394 records of check-ins. Each check-in record includes the information of user id, location, and check-in timestamp.
(2) **Melbourne Pedestrian Count Data:** We utilize the pedestrian count data during 2015/08/01 – 2015/12/31 to represent the crowd flow movements in Melbourne. The count records are composed of information of hourly counts of the pedestrians around each station, the corresponding timestamp, and the location.
(3) **Chicago Bike-Sharing Usage Data:** For Chicago, we consider the bike-sharing usage data[1] during 2016/04/01 – 2016/09/31 as the input crowd mobility data for `M-STAP`. A completed trip record in the data includes the key information of the start and end stations' ids, the pick-up and drop-off locations of the stations, and the corresponding timestamps.

Given above datasets, we show the heatmaps of the crowd flow distributions in NYC, Melbourne, and Chicago in Fig. 2.

• *Anomaly Event Data*: We have further collected several anomaly event datasets from NYC[2], Melbourne, and Chicago:

[1]https://www.divvybikes.com/
[2]https://opendata.cityofnewyork.us/



(a) Check-ins in NYC. (b) Pedestrian counts in Melbourne. (c) Bike-sharing usage in Chicago.

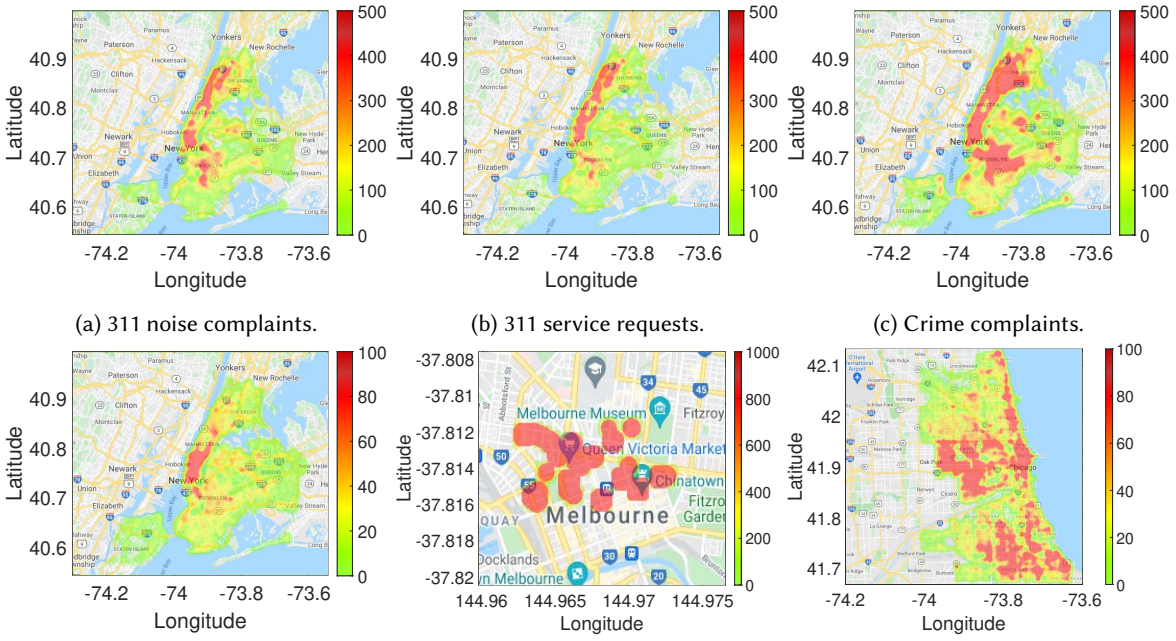
Fig. 2. Visualization of crowd flow data in NYC during 2012/07/01 — 2012/08/22, Melbourne during 2015/08/01 -– 2015/12/31, and Chicago during 2016/04/01 — 2016/09/31.

(1) **NYC Motor Vehicle Collision Data:** This dataset covers motor vehicle collision[3] in NYC during 2012/07/01 – 2012/08/22, including 24,468 collision records. The information of each collision record mainly includes location, collision time, vehicle types, and the casualties.
(2) **NYC Crime Complaint Data:** This data includes all valid felony, misdemeanor, and violation crimes reported to the NYC Police Department. The data are chosen in 2012/04/02 – 2012/08/22 and composed of 204,946 records. Each record includes its location coordinates, timestamp, and the offense types of the complaint.
(3) **NYC 311 Service Request Data:** Service request data[4] are the 311 calls or the service requests of NYC. There are 82,608 request records in the data during 2012/04/02 – 2012/08/22. Each of the logged request contains the service request location, request timestamp, and the requested service type.
(4) **NYC 311 Noise Complaint Data:** 311 Noise complaints[5] of NYC during 2012/04/02 – 2012/08/22, including 95,217 records. The detail information of each record are complaint location and timestamp.
(5) **Melbourne On-street Heavy Parking Data:** The chosen parking records are in time period of 2015/08/01 – 2015/12/31, including 9,171,718 parking records[6] in Melbourne. The heavy parking records mainly log the formation of location, the parking timestamp, and the street name. Since parking availability has been a significant issue for many metropolitan cities, we aim at evaluating the heavy parking status to represent the potential urban anomalous events.
(6) **Chicago Crime Data:** Here we use the crime data[7] in the City of Chicago, collected in 2016/04/01 – 2016/09/31. Each crime record is composed of timestamp, type, latitude and longitude.

2.2 Spatial and Temporal Data Analysis

• *Spatial Analysis*: To illustrate the data distribution of crowd flows and anomaly events, we further visualize the anomaly event data in Fig. 3 of all the three cities. Taking NYC as an example, from Figs. 3a, 3b, 3c, and 3d, we can see that anomaly events are concentrated in the downtown areas (such as Manhattan Island and Brooklyn) of NYC. Similar spatial distribution applies to the crowd flows of NYC as shown in Fig. 2a. These approximate distribution

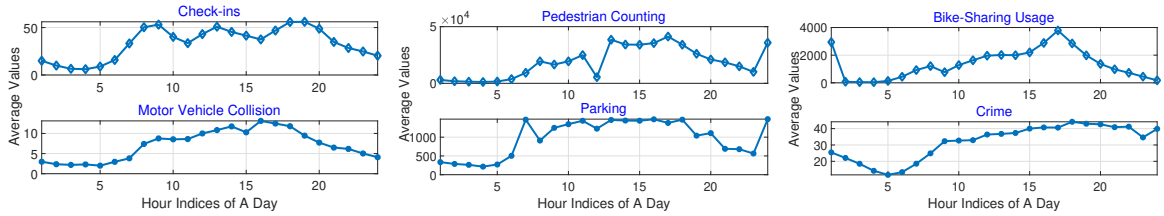
[3]https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
[4]https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
[5]https://data.cityofnewyork.us/Social-Services/311-Noise-Complaints/p5f6-bkga
[6]https://data.melbourne.vic.gov.au/Transport/On-street-Car-Parking-Sensor-Data-2015/apua-t2tb
[7]https://data.world/publicsafety/chicago-crime



(a) 311 noise complaints. (b) 311 service requests. (c) Crime complaints.

(d) Motor vehicle collisions. (e) Heavy parking events. (f) Crime events.

Fig. 3. Spatial distributions of (a) 311 noise complaints; (b) 311 service requests; (c) crime complaints in NYC during 2012/04/02 – 2012/08/22; (d) motor vehicle collisions in NYC during 2012/07/01 – 2012/08/22; (e) parking events in Melbourne during 2015/08/01 – 2015/12/31; and (f) crime events in Chicago during 2016/04/01 – 2016/09/31.



(a) Hourly check-ins and motor vehicle collisions in NYC. (b) Hourly pedestrian counts and parking events in Melbourne. (c) Hourly bike-sharing events and crime events in Chicago.

Fig. 4. Temporal distributions of (a) check-ins and motor vehicle collision data in NYC during 2012/07/01 – 2012/08/22, (b) pedestrian counts and parking events in Melbourne during 2015/08/01 – 2015/12/31, and (c) bike-sharing usage and crime events in Chicago during 2016/04/01 – 2016/09/31.

patterns of anomaly events and crowd flows demonstrate the strong correlation between the occurrence of anomaly events and the moving of crowd flows. Anomaly events are more likely to occur in the crowded regions.

• *Temporal Analysis*: Fig. 4 shows the average hourly dynamic within a day of crowd flows and anomaly events of NYC, Melbourne, and Chicago. Generally, check-ins and anomaly events of motor vehicle collision in NYC are shown to experience burst at daytime and drops at night. The bike-sharing usage and crime events in Chicago generally increase after 9am, and may start to decrease after 7pm. The pedestrian counts and parking events in Melbourne also follow the similar trend, which increase at around 8am and decrease at around 7pm. The

Table 3. Important symbols and definitions in M-STAP.

Symbols	Definitions	Symbols	Definitions
l	Length of time steps of the inputs.	t	Index of time step to predict.
$\boldsymbol{X}_{\text{in}}$	Anomaly event and crowd flow heatmap input of TDSAM, with length of l.	$\boldsymbol{X}_{\text{hout}}$	The anomaly event and crowd flow spatial feature maps output of MHSAM.
N	Number of attention heads.	$\boldsymbol{O}^n$	Output of the n^{th} attention head, $n \in N$.
$\boldsymbol{X}_{\text{cout}}$	l time steps of anomaly event spatial feature maps from the Convolution Operation in TDSAM.	$\boldsymbol{X}_{\text{sout}}$	Anomaly event output of TDSAM with length of l time steps.
K'	l time steps of output hidden states of CFSAM.	$\boldsymbol{E}^H, \boldsymbol{E}^W$	The relative height and width position matrices between each two regions of the input heatmaps.
$H \times W$	Size of input heatmaps in each time step of $\boldsymbol{X}_{\text{in}}$.	$H' \times W'$	Size of the predicted heatmap.
F_{in}	Size of input filter.	F_{out}	Size of output filter in each attention head in MHSAM.
p	Index of time steps in $\{t-l, \ldots, t-1\}$.	A_p^n	Attention score matrix evaluating the spatial correlation between each two regions in MHSAM.
D_k, D_v	Depth of the query/key projections and value projection in MHSAM.	$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Query, key and value projections in MHSAM.
$\boldsymbol{X}_{\text{cf}}$	l time steps of crowd flow feature maps from MHSAM.	$\boldsymbol{X}_{\text{out}}$	The anomaly event prediction result.

temporal similarity between crowd flows and anomaly events enables us to predict the occurrence of anomaly event by analyzing the correlation between crowd flows and anomaly events.

3 M-STAP – MULTI-HEAD SPATIO-TEMPORAL ATTENTION FUSION DESGINS

We first present the problem formulation of M-STAP and the overview of architecture in Sec. 3.1. After that, we respectively present our detailed designs on Two-Dimensional Spatial Attention Mechanism (TDSAM) in Sec. 3.2, Crowd Flow Spatial Attention Mechanism (CFSAM) in Sec. 3.3, and Temporal Attention Mechanism (TAM) in Sec. 3.4. We summarize the important symbols and their definitions in Tab. 3.

3.1 Problem Formulation and Architecture Overview

In this study, we first partition the city into a rectangular grid map with the shape of (H, W), where H and W are the number of grids/regions of the height (latitude) and width (longitude) of the grid map. Given historical anomaly event and crowd flow heatmaps $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{l \times 2 \times H \times W}$ during the recent l time steps $\{t-l, \ldots, t-1\}$, M-STAP aims at predicting the citywide anomaly event heatmap $\boldsymbol{X}_{\text{out}} \in \mathbb{R}^{H' \times W'}$ in the time step t. H' and W' are the number of regions of the height and width of the predicted heatmap.

As shown in Fig. 5, M-STAP is composed of three major components:

(1) *Two-Dimensional Spatial Attention Mechanism* (TDSAM): Correlations exist across regions with various distances from each other. This component extracts the spatial feature maps from the anomaly event heatmaps and the crowd flow heatmaps of the l historical time steps. Specially, there are two sub-components in TDSAM, *i.e.*, a Multi-Head Self-attention Mechanism (MHSAM) and a Conventional Operation (Conv2D). Instead of merely considering the local correlations among neighborhood regions, MHSAM evaluates the *pair-wise* (global) spatial correlations among all the input regions in each time step. In the meantime, the Convolution Operation extracts the local spatial correlations among the nearby regions. We concatenate the extracted global spatial feature maps from MHSAM with the local spatial feature maps from the Convolution Operation, and form the final spatial feature maps of TDSAM.

(2) *Crowd Flow Spatial Attention Mechanism* (CFSAM): Crowd flows from different regions can impact or are correlated with the occurrences of anomaly events in varying scales or degrees. We design CFSAM as an attention mechanism to capture the most relevant regions of crowd flows for the anomaly events in each time step.
(3) *Temporal Attention Mechanism* (TAM): The occurrences of anomaly events that are close in time share similar contextual information. To model above, we design a temporal attention mechanism based on Long Short-Term Memory (LSTM) to adaptively select the anomaly events of the most relevant historical time steps for anomaly event prediction.

3.2 TDSAM – Two-dimensional Spatial Attention Mechanism

Within a period of time, the anomaly events and crowd flows in nearby regions interact closely as shown in Fig. 1. Capturing the spatial correlations between different regions would help to understand the spatial dynamic trends of anomaly events and crowd flows. However, merely considering the correlations among nearby regions might ignore the information from distant regions. These derive the needs to capture the spatial correlations between the anomaly events, and between the anomaly events and crowd flows across both the nearby and distant regions.

In this section, we propose TDSAM to evaluate the near and distant spatial correlations among regions. The details of two sub-components in TDSAM, MHSAM (labeled as TDSAM-(a) and TDSAM-(b)) and Convolution Operation (labeled as TDSAM-(c)), are shown in Fig. 6. In particular, MHSAM extracts the global correlations among all regions and Convolution Operation captures the local spatial correlations among nearby regions. In TDSAM, given anomaly events and crowd flows heatmaps inputs $\boldsymbol{X}_{\text{in}}$ of the l time steps, where $\boldsymbol{X}_{\text{in}} \in \mathbb{R}^{l \times F_{\text{in}} \times H \times W}$ and $F_{\text{in}} = 2$, TDSAM outputs (i) the spatial feature maps $\boldsymbol{X}_{\text{sout}} \in \mathbb{R}^{l \times H' \times W'}$ of anomaly events and (ii) the spatial feature maps $\boldsymbol{X}_{\text{cf}} \in \mathbb{R}^{l \times H' \times W'}$ of crowd flows.

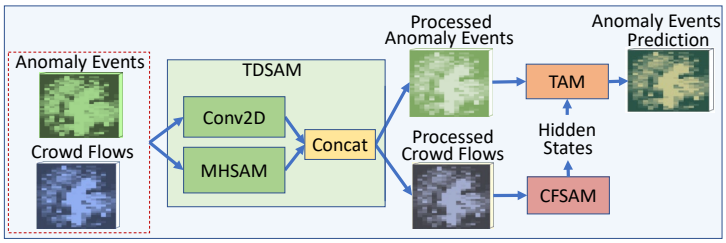


Fig. 5. Framework overview of M-STAP, including TDSAM, CFSAM, and TAM.

Instead of directly performing single feature extraction operation on the input heatmaps $\boldsymbol{X}_{\text{in}}$, MHSAM captures the global spatial features of the anomaly events and crowd flows N times in N attention heads with different data projections. Similar to the seq2seq modeling and image classification [2, 5, 27], we leverage multiple self-attention mechanisms to enhance the model learning.

As shown in Fig. 7, after feeding the input heatmaps into 2D convolution, we project the anomaly event and crowd flow heatmaps into N different projections in MHSAM by using different query, key, and value matrices with linear transformation. To obtain the spatial features of a specific region in each attention head, MHSAM calculates the weighted correlations across this region and all other neighboring regions. The height and width (number of regions) of the output feature maps in each attention head are the same as the input heatmaps. We then concatenate the extracted feature map outputs from the N attention heads and obtain the output feature maps of MHSAM. Considering the locality nature of the citywide crowd flows and anomaly events, we then leverage the Convolution Operation (Conv2D) to capture the local spatial feature maps of the crowd flows and anomaly event heatmaps. Then, we concatenate the global and local spatial feature maps in each time step, which jointly capture the spatial feature maps of both the anomaly events and crowd flows. We further apply a MaxPooling operation on the crowd flow feature maps to generate the spatial features of crowd flows in each time step with a

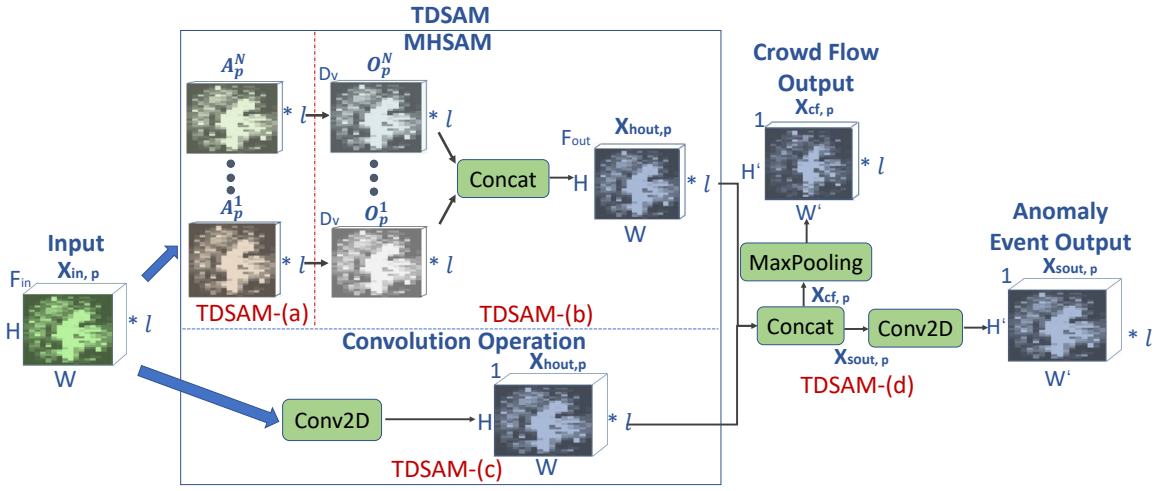


Fig. 6. Illustration of Two-Dimensional Spatial Attention Mechanism (TDSAM).

modified shape of (H', W'). A 2D convolution operation is used to generate the final citywide anomaly event feature maps of TDSAM using the concatenated heatmaps.

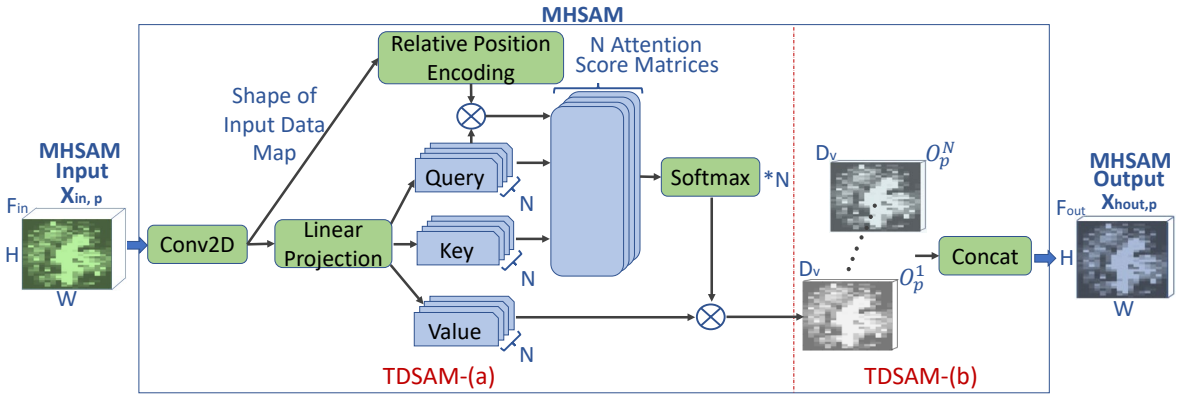


Fig. 7. Illustration of Multi-Head Self-attention Mechanism (MHSAM), which takes in the crowd flows and anomaly event data maps in time step p as input.

We present the core designs with respect to different components in TDSAM as follows.

(a) ***Attention Score for Region Correlation***. The degree of spatial correlation varies across different city regions. In particular, the geographical distance across different city regions is the key factor affecting the co-occurrences of the anomaly events. Therefore, in this component we further quantify the correlations between each two regions by evaluating the corresponding correlation attention scores. Fig. 7 further details the structure of MHSAM.

Specifically, let $\boldsymbol{X}_{in} \in \mathbb{R}^{l \times F_{\text{in}} \times H \times W}$ be the input of TDSAM module in the time steps $\{t-l, \ldots, t-1\}$. Given F_{in} channels of image-like inputs of the shape (H, W) in the time step $p \in \{t-l, \ldots, t-1\}$, we first feed the inputs

into a 2D convolution, and then flatten the inputs into a matrix $\boldsymbol{X}_{\text{in, p}}$ with the shape of (HW, F_{in}). In this study, the image-like inputs in the time step p are the crowd flow and anomaly event heatmaps. Then MHSAM linearly projects $\boldsymbol{X}_{\text{in},p}$ into different query, key, and value projections in each attention head by

$$\boldsymbol{Q}_p^n = \boldsymbol{X}_{\text{in},p}\boldsymbol{W}_q^n, \quad \boldsymbol{K}_p^n = \boldsymbol{X}_{\text{in},p}\boldsymbol{W}_k^n, \quad \boldsymbol{V}_p^n = \boldsymbol{X}_{\text{in},p}\boldsymbol{W}_v^n, \tag{1}$$

where $\boldsymbol{Q}_p^n$, $\boldsymbol{K}_p^n$ and $\boldsymbol{V}_p^n$ are the query, key, and value projections of $\boldsymbol{X}_{\text{in, }p}$ in the time step p of the attention head $n \in N$. $\boldsymbol{W}_q^n \in \mathbb{R}^{F_{\text{in}} \times D_k}$, $\boldsymbol{W}_k^n \in \mathbb{R}^{F_{\text{in}} \times D_k}$ and $\boldsymbol{W}_v^n \in \mathbb{R}^{F_{\text{in}} \times D_v}$ are the corresponding distinct learnable linear transformations matrices in the attention head n. We use the scalars $D_v \in \mathbb{R}$ and $D_k \in \mathbb{R}$ of the same value in all the attention heads.

Based on above, the global correlation attention scores of all the regions in time step p of the attention head n, denoted as $\boldsymbol{A}_p^n \in \mathbb{R}^{HW \times HW}$, are then given by

$$\boldsymbol{A}_p^n = \frac{(\boldsymbol{X}_{\text{in, }p}\boldsymbol{W}_q^n)(\boldsymbol{X}_{\text{in, }p}\boldsymbol{W}_k^n)^\top}{\sqrt{D_k}}. \tag{2}$$

In particular, the attention score $\boldsymbol{A}_{p,(i,j)}^n$ between the region $i \in \{1, 2, ..., HW\}$ and the region $j \in \{1, 2, ..., HW\}$ of $\boldsymbol{X}_{\text{in},p}$ in time step p of the attention head n can be formulated by Eq. (3) and scaled by $\frac{1}{\sqrt{D_k}}$ as

$$\boldsymbol{A}_{p,(i,j)}^n = \frac{\boldsymbol{Q}_{p,i}^n \left(\boldsymbol{K}_{p,j}^n\right)^\top}{\sqrt{D_k}}, \tag{3}$$

where $\boldsymbol{Q}_{p,i}^n \in \mathbb{R}^{1 \times D_k}$ is the query vector of the region i, and $\boldsymbol{K}_{p,j}^n \in \mathbb{R}^{1 \times D_k}$ is the key vector of the region j in the time step p of the attention head n.

The above attention score between each pair of regions is equivalent to reordering [5]. In other words, the correlation between regions which is represented by the attention score becomes independent from the location of anomaly events. This, however, can be problematic in our case, since the 2D position information (2D geological distance) between regions should be an important dimension in quantifying the spatial correlations across different city regions. In this study, we consider the 2D position information between pairs of regions by integrating their relative height and width information as proposed in [2], formulating the two-dimensional self-attention in each attention head. Specifically, given the height indices and width indices of the region i, (H_i, W_i), and another region j, (H_j, W_j), in the 2D data map, the relative height $\boldsymbol{E}_{i,j}^H \in \mathbb{R}^{1 \times D_k}$ and relative width $\boldsymbol{E}_{i,j}^W \in \mathbb{R}^{1 \times D_k}$ positions between these two regions are encoded and embedded as

$$\boldsymbol{E}_{i,j}^H = \text{Embedding}(H_i - H_j), \quad \boldsymbol{E}_{i,j}^W = \text{Embedding}(W_i - W_j). \tag{4}$$

With the relative position between two regions embedded, the attention score between the regions i and j, denoted as $\boldsymbol{A}_{p,(i,j)}^n \in \mathbb{R}^+$, is modified from Eq. (3) into

$$\boldsymbol{A}_{p,(i,j)}^n = \frac{\boldsymbol{Q}_{p,i}^n \left(\boldsymbol{K}_{p,j}^n\right)^\top + \boldsymbol{Q}_{p,i}^n \left(\boldsymbol{E}_{i,j}^H\right)^\top + \boldsymbol{Q}_{p,i}^n \left(\boldsymbol{E}_{i,j}^W\right)^\top}{\sqrt{D_k}}. \tag{5}$$

(b) ***Attention Head Output***. With the weighted correlations across regions, we can extract the spatial feature map of crowd flows and anomaly events in each attention head. Having the attention score matrix $\boldsymbol{A}_p^n$, query projection $\boldsymbol{Q}_p^n$, key projection $\boldsymbol{K}_p^n$ and value projection $\boldsymbol{V}_p^n$ in the time step p of the attention head n, we apply a self-attention layer [27] to map the input $\boldsymbol{X}_{\text{in},p}$ with the shape of (HW, F_{in}) to the feature map with the shape of

(HW, D_v) by

$$\boldsymbol{O}_p^n = \texttt{Softmax}\left(\frac{\boldsymbol{Q}_p^n \left(K_p^n\right)^\top + \boldsymbol{Q}_p^n \left(\boldsymbol{E}^H\right)^\top + \boldsymbol{Q}_p^n \left(\boldsymbol{E}^W\right)^\top}{\sqrt{D_k}}\right)\left(\boldsymbol{V}_p^n\right), \tag{6}$$

where $\boldsymbol{E}^H \in \mathbb{R}^{HW \times HW \times D_k}$ and $\boldsymbol{E}^W \in \mathbb{R}^{HW \times HW \times D_k}$ are the relative height and relative width positions between each two regions of $\boldsymbol{X}_{\text{in},\,p}$.

As mentioned above, MHSAM captures different feature maps of crowd flows and anomaly events in each attention head by diverse projections. To further utilize these feature maps with different focus, we concatenate the output feature maps from the N attention heads. In particular, the outputs of N attention heads in the time step p are concatenated and transformed into dimension F_{out} to finally output $\boldsymbol{X}_{\text{hout},\,p}$ of MHSAM. The concatenation and projection operations are formulated as

$$\boldsymbol{X}_{\text{hout},\,p} = \texttt{Concat}\left[\boldsymbol{O}_p^1, \boldsymbol{O}_p^2, ..., \boldsymbol{O}_p^N\right] \boldsymbol{W}^o + \boldsymbol{b}_{\text{out}}, \tag{7}$$

where $\boldsymbol{W}^o \in \mathbb{R}^{F_{\text{out}} \times F_{\text{out}}}$ is the learnable linear transformation, and $\boldsymbol{b}_{\text{out}} \in \mathbb{R}^{F_{\text{out}}}$ is a bias term.

(c) ***Convolution Operation***. To further take advantage of local feature processing via convolution operation, we feed the same inputs $\boldsymbol{X}_{\text{in},\,p}$ of MHSAM with shape of (F_{in}, H, W) in the time step p into the traditional Convolution Operation (Conv2D). The extracted local feature map in the time step p is then denoted as $\boldsymbol{X}_{\text{cout},\,p}$.

(d) ***Attenion Feature Fusion***. Given the global spatial feature maps from MHSAM and local spatial feature maps from Convolution Operation, we further leverage both feature maps to extract the spatial features of anomaly events in each time step. We then concatenate the output $\boldsymbol{X}_{\text{cout},\,p} \in \mathbb{R}^{H \times W}$ with $\boldsymbol{X}_{\text{hout},\,p}$, *i.e.,*

$$\boldsymbol{X}_{\text{sout},\,p} = \texttt{Concat}\left[\boldsymbol{X}_{\text{cout},\,p}, \boldsymbol{X}_{\text{hout},\,p}\right]. \tag{8}$$

In this study, F_{in} and F_{out} are both set as 2. $\boldsymbol{X}_{\text{sout},\,p}$ includes the feature maps of crowd flows and anomaly events. Given above, we denote the spatial feature map of crowd flows in $\boldsymbol{X}_{\text{sout},\,p}$ in the time step p as $\boldsymbol{X}_{\text{cf},p} \in \mathbb{R}^{H \times W}$. Then we apply MaxPooling operation to modify the crowd flow heatmap in the time step p into the shape of (H', W') by

$$\boldsymbol{X}_{\text{cf},p} = \texttt{MaxPooling}(\boldsymbol{X}_{\text{cf},p}). \tag{9}$$

We further apply a 2D convolution upon $\boldsymbol{X}_{\text{sout},\,p}$ to generate the spatial feature map of anomaly events in the time step p, *i.e.,*

$$\boldsymbol{X}_{\text{sout},\,p} = \texttt{Conv2D}(\boldsymbol{X}_{\text{sout},\,p}). \tag{10}$$

The spatial features of anomaly events account for both the distribution of crowd flows and anomaly events. The shape of $\boldsymbol{X}_{\text{sout},\,p}$ now is (H', W'). The l time steps' crowd flow outputs and anomaly event outputs of TDSAM are denoted as $\boldsymbol{X}_{\text{cf}}$ and $\boldsymbol{X}_{\text{sout}}$, respectively.

3.3 CFSAM – Crowd Flow Spatial Attention Mechanism

The dynamic crowd flows can reflect the occurrences of crowd-related anomaly events. However, not all regions of crowd flows are related to the occurrences of anomaly events of a specific region. Due to the sparseness of the anomaly events and crowd flows, there are also some regions without any crowd flows and anomaly events for some time steps. Inspired by [24], we propose an attention mechanism to select the relevant regions of crowd flows for each region of anomaly events in each time step. Specifically, given the crowd flow feature maps $\boldsymbol{X}_{\text{cf}}$ during time steps $\{t-l, \ldots, t-1\}$ from TDSAM, we capture the importance of crowd flows from different regions in each time step in Crowd Flow Spatial Attention Mechanism (CFSAM). The details of CFSAM are illustrated in Fig. 8.

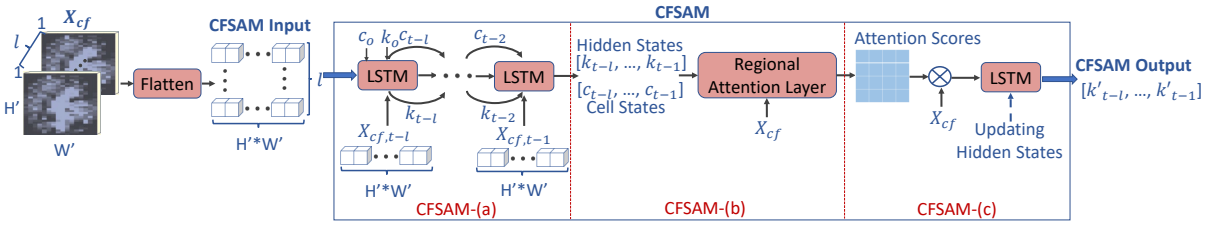


Fig. 8. Illustration of Crowd Flow Spatial Attention Mechanism (CFSAM). $\boldsymbol{c}_o$ & $\boldsymbol{k}_o$ are the initial cell state and hidden state.

(a) LSTM *for Crowd Flow Heatmap*. With the integration of the memory cell, input gate, output gate and forget gate, LSTM is capable of capturing the long-term dependencies of time series by summing the features of anomaly activities over time, and overcoming vanishing gradient problem. Inspired by above, we use LSTM in the attention mechanism to process the crowd flows from different regions and different time steps. First, we flatten the crowd flow feature maps $\boldsymbol{X}_{\text{cf}}$ into shape of $(l, H'W')$. CFSAM then maps the flattened crowd flow feature map in the time step p, $\boldsymbol{X}_{\text{cf},\,p}$, $p \in \{t-l, \ldots, t-1\}$, into $\boldsymbol{k}_p \in \mathbb{R}^m$, *i.e.,*

$$\boldsymbol{k}_p = \texttt{LSTM}(\boldsymbol{k}_{p-1}, \boldsymbol{X}_{\text{cf},\,p}), \tag{11}$$

which is the hidden state of Long Short-Term Memory (LSTM) in the time step p. Specifically, the LSTM cell in the time step p is formally given by

$$\boldsymbol{f}_p = \sigma\left(\boldsymbol{W}_f[\boldsymbol{k}_{p-1}, \boldsymbol{X}_{\text{cf},\,p}] + \boldsymbol{b}_f\right), \tag{12}$$

$$\boldsymbol{u}_p = \sigma\left(\boldsymbol{W}_u[\boldsymbol{k}_{p-1}, \boldsymbol{X}_{\text{cf},\,\text{p}}] + \boldsymbol{b}_u\right), \tag{13}$$

$$\boldsymbol{v}_p = \sigma\left(\boldsymbol{W}_v[\boldsymbol{k}_{p-1}, \boldsymbol{X}_{\text{cf},\,p}] + \boldsymbol{b}_v\right), \tag{14}$$

$$\boldsymbol{c}_p = \boldsymbol{f}_p \odot \boldsymbol{c}_{p-1} + \boldsymbol{u}_p \odot \tanh\left(\boldsymbol{w}_c[\boldsymbol{k}_{p-1}, \boldsymbol{X}_{\text{cf},\,p}] + \boldsymbol{b}_c\right), \tag{15}$$

$$\boldsymbol{k}_p = \boldsymbol{v}_p \odot \tanh(\boldsymbol{c}_p), \tag{16}$$

where $\boldsymbol{c}_p$ is the cell state of *LSTM* cell in the time step p, $\boldsymbol{W}_f \in \mathbb{R}^{m\times(m+1)}$, $\boldsymbol{W}_u \in \mathbb{R}^{m\times(m+1)}$, $\mathbf{W_v} \in \mathbb{R}^{\mathbf{m\times(m+1)}}$, $\boldsymbol{W}_c \in \mathbb{R}^{m\times(m+1)}$, $\boldsymbol{b}_f \in \mathbb{R}^m$, $\boldsymbol{b}_u \in \mathbb{R}^m$, $\boldsymbol{b}_v \in \mathbb{R}^m$, and $\boldsymbol{b}_c \in \mathbb{R}^m$ are parameters to learn. $[k_{p-1}, \boldsymbol{X}_{\text{cf},\,p}] \in \mathbb{R}^{m+1}$ is the concatenation of the hidden state in the time step $(p-1)$ and the input of the crowd flow in current time step. $\sigma(\cdot)$ and $\odot$ are the Sigmoid function and element-wise multiplication operation, respectively.

(b) *Attention for Region Crowd Flow Importance*. Given $H'W'$ regions of crowd flows of l historical time steps, we design a *regional attention layer* to measure the importance of crowd flows from different regions in each time step. The attention weight of crowd flow from the region i in the time step p is calculated by

$$e_p^i = \boldsymbol{w}_e^\top \tanh(\boldsymbol{W}_e[\boldsymbol{k}_{p-1}, \boldsymbol{c}_{p-1}] + \boldsymbol{U}_e \boldsymbol{X}_{\text{cf}}^i), \quad a_p^i = \frac{\exp(e_p^i)}{\sum_{i'=1}^{H'W'} \exp(e_p^{i'})}, \tag{17}$$

where a_p^i is the attention weight of crowd flow $\boldsymbol{X}_{\text{cf}}^i$ of the region $i \in \{1, 2, ..., H'W'\}$ in the time step p. $\boldsymbol{w}_e^\top \in \mathbb{R}$, $\boldsymbol{W}_e \in \mathbb{R}^{1\times 2m}$, and $\boldsymbol{U}_e \in \mathbb{R}$ are the parameters to learn. Note that the attention weights of all crowd flows in one time step sum to 1.

(c) *Crowd Flow Attention Weight & Hidden State Updates*. After measuring the importance of crowd flows using the attention weights, the crowd flows of all regions in the time step p are then updated by

$$\widetilde{\boldsymbol{X}}_{\text{cf},p} = \left\{a_p^1 \boldsymbol{X}_{\text{cf},p}^1, \quad a_p^2 \boldsymbol{X}_{\text{cf},p}^2, \quad \ldots, \quad a_p^{H'W'} \boldsymbol{X}_{\text{cf},p}^{H'W'}\right\}. \tag{18}$$

The value of the crowd flow entries after re-weighting is denoted as

$$\widetilde{\boldsymbol{X}}_{\text{cf}} = \left\{\widetilde{\boldsymbol{X}}_{\text{cf},t-l}, \quad \widetilde{\boldsymbol{X}}_{\text{cf},t-l+1}, \quad ..., \quad \widetilde{\boldsymbol{X}}_{\text{cf},t-1}\right\}. \tag{19}$$

Given the weighted crowd flows from $H'W'$ regions of l time steps, M-STAP can particularly learn upon the most important regions of crowd flows in each time step by updating the hidden states $k \in \mathbb{R}^{l\times m}$ with the updated crowd flows. The hidden state in the time step p is updated as

$$\boldsymbol{k}'_p = \text{LSTM}\left(\boldsymbol{k}_{p-1}, \widetilde{\boldsymbol{X}}_{\text{cf},p}\right). \tag{20}$$

The hidden states during time steps $\{t-l,\ldots,t-1\}$ are denoted as $\boldsymbol{k}' = \{\boldsymbol{k}'_{t-l}, \boldsymbol{k}'_{t-l+1}, ..., \boldsymbol{k}'_{t-1}\}$.

3.4 TAM – Temporal Attention Mechanism

Given l time steps of anomaly event feature maps $\boldsymbol{X}_{\text{sout}}$ from TDSAM and l time steps of hidden states $\boldsymbol{k}'$ from CFSAM, we propose Temporal Attention Mechanism (TAM) to differentiate the contributions of anomaly events in l historical time steps on the prediction.

(a) ***Attention for Temporal Hidden States***. Given l time steps of hidden states $\boldsymbol{k}'$ which measure the importance of crowd flows from different regions in each time step, we can further weigh the influences of crowd flows from l time steps on the anomaly events in each time step. In particular, as shown in Fig. 9, the attention weight $\tilde{a}_p^q$ of the hidden state $\boldsymbol{k}'_q$ in TAM on the anomaly events in the time step p can be calculated by the regional attention layer, *i.e.,*

$$d_p^q = \boldsymbol{w}_d^\top \tanh(\boldsymbol{W}_d[\boldsymbol{h}_{p-1}, \boldsymbol{c}'_{p-1}] + \boldsymbol{U}_d\boldsymbol{k}'_q + \boldsymbol{b}_d), \quad \tilde{a}_p^q = \frac{\exp(d_p^q)}{\sum_{q'=t-l}^{t-l}\exp(d_p^{q'})}, \tag{21}$$

where $q \in \{t-l,\ldots,t-1\}$ and $p \in \{t-l,\ldots,t-1\}$. $[\boldsymbol{h}_{p-1}, \boldsymbol{c}'_{p-1}] \in \mathbb{R}^{2m}$ is the concatenation between the hidden state $\boldsymbol{h}_{p-1}$ and cell state $\boldsymbol{c}'_{p-1}$ calculated by LSTM using the heatmaps of anomaly events $\boldsymbol{X}_{\text{sout},p-1}$ in the time step $p-1$. $\boldsymbol{w}_d^\top \in \mathbb{R}$, $\boldsymbol{W}_d \in \mathbb{R}^{1\times 2m}$, $\boldsymbol{U}_d \in \mathbb{R}^{1\times m}$, and $\boldsymbol{b}_d \in \mathbb{R}$ are parameters to learn.

Then we can compute the weighted sum of the influences of all hidden states $\boldsymbol{k}'$ on the anomaly events in the time step p. The weighted sum is represented by context vector $\boldsymbol{g}_p \in \mathbb{R}^m$ at the time step p, which is calculated as

$$\boldsymbol{g}_p = \sum_{q=t-1}^{t-l} \tilde{a}_p^q \boldsymbol{k}'_q. \tag{22}$$

The context vectors during time steps $\{t-l,\ldots,t-1\}$ are given by $\boldsymbol{g} = \{g_{t-l}, g_{t-l+1}, \ldots, g_{t-1}\}$.

(b) ***Anomaly Event Heatmap & Hidden State Update***. Accounting for both the influences from the crowd flows and previous anomaly events, we update the anomaly event heatmaps $\boldsymbol{X}_{\text{sout}}$ during time steps $\{t-l,\ldots,t-1\}$ by the *linear operation, i.e.,*

$$\widetilde{\boldsymbol{X}}_{\text{sout},p}^{h,w} = \widetilde{\boldsymbol{w}}^\top \left[\boldsymbol{X}_{\text{sout},p}^{h,w}, \boldsymbol{g}_p\right] + \widetilde{\boldsymbol{b}}, \tag{23}$$

where $\widetilde{\boldsymbol{X}}_{\text{sout},p}^{h,w}$ is the updated anomaly event in grid (h,w) in the time step p, $h \in \{1,\ldots,H'\}$ and $w \in \{1,\ldots,W'\}$. $[\boldsymbol{X}_{\text{sout},p}^{h,w}, \boldsymbol{g}_p]$ is the concatenation between $\boldsymbol{X}_{\text{sout},p}^{h,w}$ and the context vector $\boldsymbol{g}_p$ in the time step p. $\widetilde{\boldsymbol{w}}^\top \in \mathbb{R}^{1\times(m+1)}$ and $\widetilde{\boldsymbol{b}} \in \mathbb{R}$ are parameters to learn.

Having the updated anomaly events of each region in each time step, we can calculate the corresponding hidden state h_p and cell state c'_p of LSTM in the time step p by

$$h_p, c'_p = \text{LSTM}\left(h_{p-1}, c'_{p-1}, \widetilde{\boldsymbol{X}}_{\text{sout},p}\right). \tag{24}$$

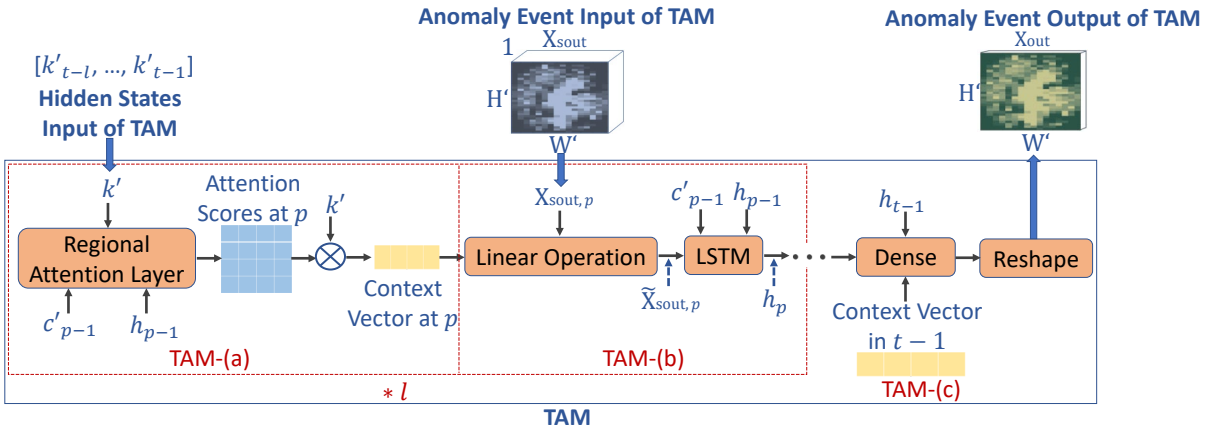


Fig. 9. Illustration of Temporal Attention Mechanism (TAM).

(c) *Final Anomaly Event Distribution Output*. In this step, we have the context vector $\boldsymbol{g}_{t-1}$ at the time step $t-1$ which aggregates the influences of both the previous l time steps of crowd flows and anomaly events. Then we calculate the anomaly event prediction of region (h, w) in the time step t with $\boldsymbol{g}_{t-1}$ and the hidden state $\boldsymbol{h}_{t-1}$ at the time step $t-1$ by the Dense and ReLU activation operations, *i.e.,*

$$\boldsymbol{X}_{\text{out},t}^{h,w} = \widetilde{\boldsymbol{w}}_b^{\top} \text{ReLU}\left(\widetilde{\boldsymbol{W}}_b[\boldsymbol{h}_{t-1}, \widetilde{\boldsymbol{g}}_{t-1}] + \widetilde{b}_b\right), \tag{25}$$

where $\widetilde{\boldsymbol{w}}_b^{\top} \in \mathbb{R}$, $\widetilde{\boldsymbol{W}}_b \in \mathbb{R}^{1\times 2m}$ and $\widetilde{b}_b \in \mathbb{R}$ are parameters to learn. Having all the predictions in the time step t, we reshape the predictions into shape of $H' \times W'$. The final predictions of all regions in the time step t are denoted as $\boldsymbol{X}_{\text{out}} \in \mathbb{R}^{H'\times W'}$.

4 EXPERIMENTAL STUDIES

We first present the experimental settings in Sec. 4.1, and then we show the evaluation results in Sec. 4.2.

4.1 Experimental Settings

We compare our M-STAP with the following baselines.

(1) *Historical Average* (HA): The anomaly events of each grid in the time step t are predicted as the average of the data in the same time step of the days during time steps $\{t-l, \ldots, t-1\}$.
(2) *Gaussian Process* (GP): In the Gaussian Process (GP) time series model [11], a total of l historical time steps of anomaly event data during time steps $\{1, \ldots, t-1\}$ are used to predict the anomaly events of each region in the time step t.
(3) *Recurrent Neural Networks* (RNN): We implement RNN in this study to predict the time series of the anomaly events in the time step t, and the number of historical time steps l is set as 4.
(4) *Long Short-Term Memory* (LSTM): LSTM takes in the l historical time steps of anomaly event heatmaps and predicts the anomaly events in the time step t.
(5) *Gated Recurrent Unit* (GRU): We feed the l historical time steps of anomaly event data into GRU to predict the anomaly events in the incoming time step t.
(6) *Convolutional LSTM Network* (ConvLSTM): Given crowd flows and anomaly events heatmaps X_{in} in time steps $\{t-l, \ldots, t-1\}$, ConvLSTM predicts the anomaly event heatmaps in the time step t.

(7) TPA-LSTM [26]: which leverages attention-recurrent neural network to predict the anomaly events heatmap with the shape of $H' \times W'$ by using l historical time steps of anomaly event heatmaps with the same shape as the predicted heatmap.
(8) CHAT [15]: We adapt the Cross-Interaction Hierarchical Attention (CHAT) network to predict the frequency of multi-region anomaly event. We use l historical time steps of anomaly events with the shape of $H' \times W'$ to predict the next time step of anomaly event heatmap with the shape of $H' \times W'$.
(9) MTGNN [32]: In Multivariate Time-series forecasting with Graph Neural Networks (MTGNN), we use the l historical time steps of anomaly event and crowd flow data with the shape of $H \times W$ to predict one following time step of anomaly event heatmap with the shape of $H' \times W'$.
(10) STResNet [37]: We consider that historical anomaly event and crowd flow heatmaps with the shape of $H \times W$ are applied to predict one following time step of the anomaly event heatmap with the shape of $H' \times W'$. The lengths of closeness, period, trend sequences in STResNet are all set as 4.

Unless otherwise stated, we use the following parameter settings by default. For the data of NYC, $H \times W$, $H' \times W'$, l, m, N, D_k, D_v, learning rate, batch size, and training epochs are set as 24×24, 16×16, 4, 16, 2, 2, 1, 0.0002, 128 and 4,000, respectively. For the data of Chicago, $H \times W$, $H' \times W'$, l, m, N, D_k, D_v, learning rate, batch size, and training epochs are set as 16×9, 16×9, 4, 5, 2, 2, 1, 0.0002, 128 and 1,000. For the data of Melbourne, $H \times W$, $H' \times W'$, l, m, N, D_k, D_v, learning rate, batch size, and training epochs are set as 26×26, 26×26, 4 16, 2, 2, 1, 0.0002, 256 and 2,000. For datasets of 311 noise complaint, 311 service request, and crime complaint in NYC, as well as both Chicago and Melbourne, we leave out the last 40 days for validation and testing, *i.e.,* the first 20 days are for validation and the next 20 days for testing, and rest are used for training. For motor vehicle collision in NYC, we use the last 20 days for validation and testing, *i.e.,* the first 10 days are for validation and the next 10 days for testing, and the rest are used for training. Due to sparsity of the dataset, the length of one time step is 12h for NYC and Chicago. For Melbourne we adopt 1h for each time step. All the experiments are trained based on the loss of the Mean Squared Error (MSE), *i.e.,*

$$\text{MSE} = \frac{1}{H' \times W'} \times \sum_{h=1}^{H'} \sum_{w=1}^{W'} \left(\widehat{X}_{\text{out}}^{h,w} - X_{\text{out}}^{h,w} \right)^2. \tag{26}$$

Our evaluation matrices include the Mean Absolute Error (MAE), the Error Rate (ER), the Root Mean Squared Error (RMSE) and the Mean Squared Logarithmic Error (MSLE) as follows,

$$\begin{aligned}
&\text{MAE} = \frac{1}{H' \times W'} \times \sum_{h=1}^{H'} \sum_{w=1}^{W'} \left| \widehat{X}_{\text{out}}^{h,w} - X_{\text{out}}^{h,w} \right|, \quad \text{ER} = \frac{\sum_{h=1}^{H'} \sum_{w=1}^{W'} \left| \widehat{X}_{\text{out}}^{h,w} - X_{\text{out}}^{h,w} \right|}{\sum_{h=1}^{H'} \sum_{w=1}^{W'} X_{\text{out}}^{h,w}}, \\
&\text{RMSE} = \sqrt{\frac{1}{H' \times W'} \times \sum_{h=1}^{H'} \sum_{w=1}^{W'} \left(\widehat{X}_{\text{out}}^{h,w} - X_{\text{out}}^{h,w} \right)^2}, \quad \text{MSLE} = \frac{1}{H' \times W'} \times \sum_{h=1}^{H'} \sum_{w=1}^{W'} \left| \log_2 \left(\widehat{X}_{\text{out}}^{h,w} + 1 \right) - \log_2 \left(X_{\text{out}}^{h,w} + 1 \right) \right|,
\end{aligned} \tag{27}$$

where $\widehat{X}_{\text{out}}^{h,w}$ is the predicted anomaly event in region (h, w) of the time step t, $h \in \{1, \ldots, H'\}$ and $w \in \{1, \ldots, W'\}$, and $X_{\text{out}}^{h,w}$ is the corresponding ground truth. The results of all experiments are the average values of two consecutive experiments with the same parameters and environment.

We conduct experimental studies on the Google Colab[8] and a desktop of Intel i7-9700, NVIDIA GeForce RTX 2060 SUPER, 16.0 GB RAM, and Windows 10. The proposed model is implemented in Python with Tensorflow-GPU-2.3.0.

[8]https://colab.research.google.com/notebooks/intro.ipynb#recent=true

Table 4. Performance comparison on urban anomaly datasets from NYC, Melbourne, and Chicago.

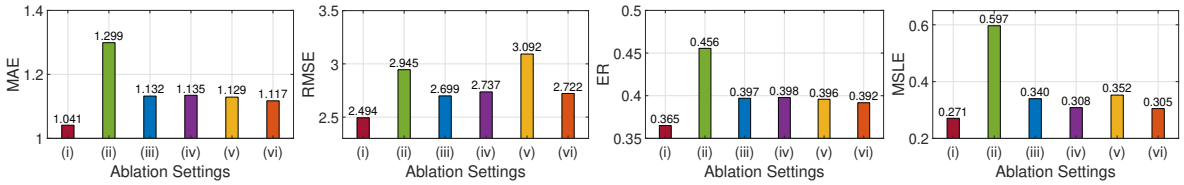
Scheme	311 Noise Complaint (NYC)				311 Service Request (NYC)				Crime Complaint (NYC)			
	MAE	RMSE	ER	MSLE	MAE	RMSE	ER	MSLE	MAE	RMSE	ER	MSLE
HA	0.870	2.919	0.678	0.407	0.750	2.751	0.669	0.334	1.301	3.660	0.456	0.310
GP	1.151	3.565	0.967	0.986	1.069	3.475	0.946	0.892	2.554	6.986	0.904	1.974
RNN	1.183	3.735	0.922	0.922	1.049	3.384	0.935	0.810	2.556	6.813	0.896	1.956
LSTM	1.140	3.623	0.889	0.821	0.968	3.160	0.863	0.702	2.468	6.681	0.866	1.633
GRU	1.175	3.650	0.917	0.852	1.041	3.297	0.928	0.790	2.496	6.730	0.875	1.760
ConvLSTM	1.184	3.780	0.924	0.734	1.030	3.441	0.919	0.633	2.657	7.044	0.932	1.555
TPA-LSTM	0.898	2.645	0.704	0.482	1.079	3.626	0.962	0.693	1.301	3.232	0.470	0.459
CHAT	0.957	2.926	0.746	0.376	0.864	2.825	0.771	0.320	1.361	3.577	0.477	0.664
MTGNN	0.955	2.828	0.775	0.222	0.832	2.586	0.739	0.203	1.375	4.007	0.481	0.230
STResNet	0.912	2.540	0.711	0.432	0.841	2.485	0.750	0.557	1.397	2.966	0.490	0.608
M-STAP	0.818	2.263	0.638	0.383	0.775	2.332	0.691	0.376	1.041	2.494	0.365	0.271

Scheme	Motor Vehicle Collision (NYC)				Parking Event (Melbourne)				Crime Event (Chicago)			
	MAE	RMSE	ER	MSLE	MAE	RMSE	ER	MSLE	MAE	RMSE	ER	MSLE
HA	0.859	2.243	0.961	0.946	1.711	12.879	0.481	0.121	1.251	2.259	0.550	0.460
GP	0.837	2.322	0.926	0.804	3.860	29.841	0.989	1.380	1.689	3.263	0.733	1.131
RNN	0.855	2.094	0.956	0.660	1.732	13.479	0.487	0.362	1.252	2.226	0.551	0.740
LSTM	0.633	1.597	0.708	0.421	1.941	15.333	0.546	0.294	1.518	2.853	0.668	1.025
GRU	0.766	1.901	0.857	0.574	1.889	15.178	0.531	0.267	1.337	2.436	0.588	0.778
ConvLSTM	0.844	2.088	0.944	0.562	2.615	12.138	0.739	1.381	1.171	2.005	0.515	0.434
TPA-LSTM	0.545	1.316	0.627	0.375	1.400	11.580	0.395	0.212	1.345	2.377	0.590	0.522
CHAT	0.648	1.611	0.745	0.376	2.007	15.324	0.567	0.228	1.173	2.132	0.525	0.453
MTGNN	0.588	1.492	0.678	0.201	1.832	11.725	0.517	0.159	1.062	1.992	0.470	0.262
STResNet	0.613	1.502	0.705	0.371	1.235	9.159	0.347	0.063	1.525	2.987	0.671	1.307
M-STAP	0.549	1.314	0.615	0.280	0.579	5.775	0.248	0.041	0.932	1.701	0.410	0.294

4.2 Evaluation Results

• **Overall Results**: Tab. 4 demonstrates the experiment results of predicting anomaly events of NYC, Melbourne and Chicago using our proposed method M-STAP and other baselines. Compared with other baselines, M-STAP demonstrates the following average improvements in all metrics considered per dataset: (i) on average 28.07% when predicting 311 noise complaints, 21.56% for 311 service requests, 55.17% for crime complaints, 30.63% for motor vehicle collision events in NYC; (ii) on average 69.64% error reduction for parking events in Melbourne; and (iii) on average 46.46% error reduction for crime events in Chicago.

Compared with M-STAP, RNN, LSTM, and GRU only capture the temporal correlations and are not able to capture the spatial correlations. ConvLSTM is a combination of convolutional neural network (CNN) and LSTM, and it fails to characterize the different contributions of different context features and time steps. HA outperforms baselines like GP and RNN/LSTM/GRU in predicting most of datasets. It is mainly because of the sparsity within the anomaly datasets where the anomaly events mostly happen at limited locations and time periods. For instance, the heavy parking events in Melbourne happen at less than 16% of the time steps at all city regions. Therefore, taking the average of historical data does not generate predictions far from ground-truths for the HA approach, while the conventional deep learning approaches may be prone to the dynamic and sparse anomaly events. GP may



(a) MAE. (b) RMSE. (c) ER. (d) MSLE.

Fig. 10. Performance of crime event prediction for NYC using different ablation settings.

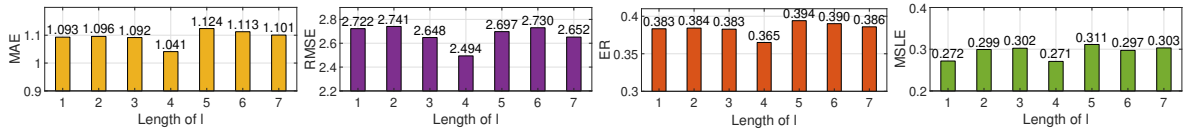
suffer from the complex dynamics and spatio-temporal variations within the anomalies and hence cannot achieve satisfactory results. Different from M-STAP, TPA-LSTM and CHAT do not consider the spatial correlations between time series of different regions. MTGNN and STResNet merely analyze the spatial correlations among neighboring local regions. Therefore, these approaches may not achieve high accuracy.

M-STAP provides higher accuracy thanks to its more comprehensive model integration with the spatio-temporal characteristics of the crowd flows (check-ins and other mobility data) and anomaly events. The results show that our proposed method has overall higher prediction accuracy compared with other baselines, demonstrating the effectiveness of M-STAP in multi-region anomaly events prediction problem.

• **Ablation Studies:** We compare the base design of (i) M-STAP, and different ablation settings, *i.e.,* (ii) M-STAP without CFSAM+TAM, (iii) M-STAP without TDSAM, (iv) M-STAP without Convolution Operation in TDSAM, (v) M-STAP without MHSAM in TDSAM, and (vi) M-STAP where the LSTM is replaced by GRU. Specifically, we use the crime complaint data of NYC as example to investigate the performance of each component of the proposed model. All the parameters are the same as the ones of predicting crime complaint data of NYC by M-STAP as mentioned above. As shown in Fig. 10, M-STAP achieves the highest accuracy, and on average improves 10.44% in MAE, 12.15% in RMSE, 10.50% in ER, and 35.89% in MSLE, when compared with the remaining five settings. The improvements from CFSAM+TAM and TDSAM are more significant by comparing the prediction result of (i) M-STAP against (ii) M-STAP without CFSAM+TAM, and (iii) M-STAP without TDSAM. The prediction results show that TDSAM is important for measuring the spatial correlations of each region of the anomaly events, while CFSAM+TAM demonstrates effectiveness in evaluating the correlation between occurrence of anomaly events and the movement of crowd flows, and capturing the most important historical time steps for prediction. In addition, leveraging both TDSAM and Convolution Operation benefits the integration of the spatial correlations of anomaly events. The occurrences of anomaly events in a specific region are correlated with the ones from different time periods. By applying LSTM in CFSAM+TAM, M-STAP is capable of capturing both the short-term and long-term dependencies of anomaly events than using GRU.

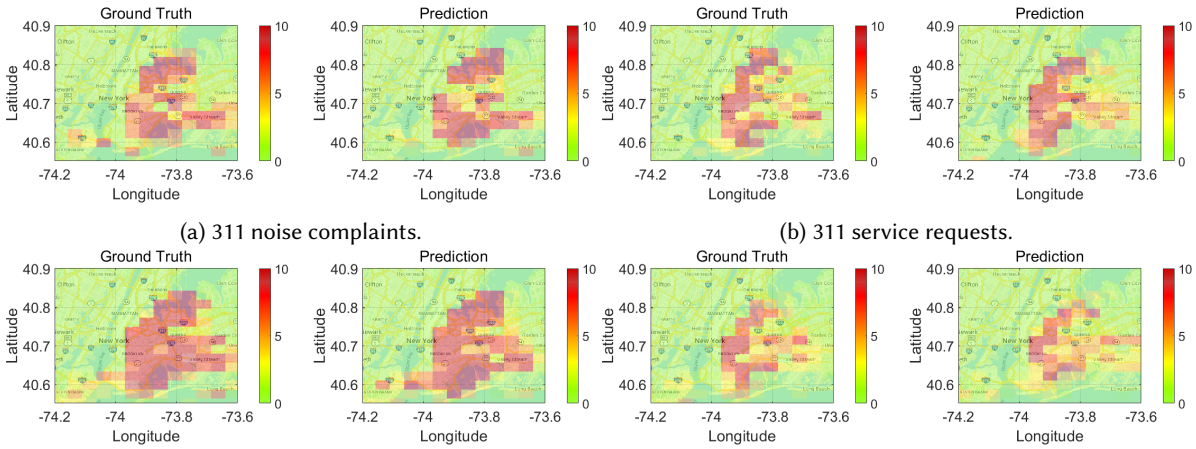
• **Sensitivity Studies**: To evaluate the number of time steps of the inputs, denoted as l, we predict the noise complaints with our proposed model with l set as 1 to 7. As shown in Fig. 11, our proposed model performs the best with l set as 4, indicating that utilizing the historical two days' crowd flows and 311 noise complaint data to predict the anomaly event in the next 12 hours gains the highest accuracy. In this study, we can see that the predictions of 311 noise complaints, 311 service requests, crime complaints, and motor vehicle collision events achieve the highest accuracy when l is set as 4.

• **Visualization**: Taking NYC as an example, Fig. 12 further illustrates the ground-truth and predicted heatmaps of 311 noise complaints, 311 service requests, crime complaints, and motor vehicle collisions in NYC during one selected time step in the testing data. In the ground-truth and predicted heatmaps of the same anomaly event, the warmer colors in a region indicate the larger number of anomaly events. The dynamic spatial and temporal dependencies of anomaly events make it hard to predict the frequencies of anomaly event of each interacted regions. M-STAP considers both factors and achieves high prediction result of the multi-region anomaly event prediction. We can observe from the heatmaps that the spatial distributions and the frequencies of each anomaly



(a) MAE. (b) RMSE. (c) ER. (d) MSLE.

Fig. 11. Prediction results of crime complaint anomaly events of NYC using different length of l.



(a) 311 noise complaints. (b) 311 service requests.

(c) Crime complaints. (d) Motor vehicle collisions.

Fig. 12. The heatmaps of the ground truths and prediction results of: (a) 311 noise complaints; (b) 311 service requests; (c) crime complaints; and (d) motor vehicle collisions of NYC of one selected time step in the testing data.

event are predicted accurately by M-STAP. The accurate multi-region anomaly event prediction results can benefit the applications such as urban emergency prediction, and subsequent crowd flow redistribution and management.

5 RELATED WORK

We briefly discuss the related work as follows. In particular, we first review the urban check-in and traffic data analysis, and then review the urban anomaly data analytics in two categories: *detection* and *prediction*.

• *Urban Check-in & Traffic Data*: Location-based social network platforms such as Foursquare [18], Twitter and DenseGPS [10] provide plenty of check-in data which can characterize the time and locations visited by users, enabling various kinds of urban applications and services [22, 23]. By understanding the correlations between trajectories, locations and users, Zheng *et al.* pioneered a location-based social network framework for friend and POI recommendation [1, 40]. Cao *et al.* [3] and Jiang *et al.* [16] respectively presented large-scale analysis of point-of-interest (POI) revisitation patterns and regions-of-interest (ROI) to model the periodic behavior of human mobility. Chen *et al.* recently identified the similarities and differences of visitations and check-ins [4], which is essential for enhancing human mobility modeling. Besides above, various mobility modeling and system designs have been investigated, including decentralized and personalized designs [8], probabilistic activity [20], and Bayesian mixture modeling [28]. Social and business values have been further derived by analyzing the urban human mobility data. Yang *et al.* studied the commercial activeness prediction based on urban big data including user check-ins [33]. Using check-in data as auxiliary information, Fan *et al.* proposed an approach for personalized context-aware collaborative online activity prediction [7]. Ruan *et al.* studied the dynamic public resource allocation on the real-world crowd flow data in a theme park in Beijing [25].

On the other hand, understanding the traffic patterns has become increasingly important for urban management and response. Many research efforts have focused on demand prediction of taxi [34] and bike-sharing [21] and estimation of travel time [29]. Wu *et al.* proposed the graph neural network for multivariate time series forecasting which was validated in traffic condition prediction [32].

Different from the above studies, we focus on leveraging the ubiquitously harvested urban mobility data for accurate anomaly prediction. In particular, we have studied the generality and adaptivity of M-STAP under different crowd flows and mobility datasets, including Foursquare check-in, bike-sharing usage, and pedestrian counting data, and leverage each as the input crowd flows serving as an important auxiliary information within our novel attention-based deep learning approach. Such a novel fusion and integration enables accurate and robust urban anomaly prediction and has been experimentally validated in multiple different anomaly datasets.

• *Anomaly Detection*: Detecting the urban anomaly events is an essential task for many city governors and planning agencies. It has been received wide attentions as it enables or boosts many important applications such as travel time estimation of multiple transportation systems [9] and exploration of anomalous regions with long-term poor traffic situations [19]. Various approaches have emerged on the detection of anomaly events. Witayangkurn *et al.* proposed a hidden markov model based on GPS data from mobile phones [30]. Different sources and domains of spatio-temporal data have been further utilized for anomaly detection [36, 41]. Zhang *et al.* developed a decomposition approach to detect urban anomaly [38]. Different from these works, we focus on *forecasting* the distributions of urban anomaly events based on historical urban data fusion and novel multi-head spatio-temporal attention designs, which will enable more proactive urban applications and predictive modeling of city patterns.

• *Anomaly Prediction*: Anomaly prediction aims at forecasting the timestamps/periods and locations of the incoming anomaly events, and can generally be categorized into the following two folds: the event *classification* and *regression*. *(a)* In classification studies, Wu *et al.* [31] designed an approach on forecasting which category of urban anomaly events will happen in each region of the city. Furthermore, Huang *et al.* proposed crowdsoursing approach to predict frequencies of anomaly events [14]. *(b)* In regression problems, Zhao *et al.* modeled the spatio-temporal correlations for crime prediction [39]. Taking in the city anomalies as input, Huang *et al.* proposed bidirectional LSTM with cross-interaction hierarchical attention for prediction of urban anomaly categories [15], where they use an interaction attention and a temporal attention to model the correlations between different regions and time periods. However, it does not take into account the effect of surrounding crowd dynamics on the occurrence of anomaly event, which is an important aspect for predicting the distributions of anomaly events [17]. Based on multi-head spatial and temporal attention, our M-STAP focuses on a more challenging task on predicting the multi-region distributions of anomaly events given historical anomalies and crowd flow distributions, and our experimental studies have validated its better performance than other state-of-the-arts.

Different from above work, we propose utilizing the crowd flow as an *auxiliary* information for enhanced anomaly event prediction. While some recent studies have investigated deep learning for crime prediction [39] and regional traffic accident prediction [35], these works often focus on the short-term impact of historical events alone, but may not fully consider the long-term contribution and the relative importance of each context feature to the target series in the same time slot. Our M-STAP falls in this category of studies, but it develops a novel multi-head attention mechanism to further differentiate the dynamic contributions of different spatial, temporal as well as contextual features. Our experimental studies have further validated the accuracy and robustness of M-STAP compared with the baseline algorithms.

6 CONCLUSION

In the study, we propose M-STAP, a Multi-head Spatio-Temporal Attention Prediction approach to address the multi-region urban anomaly events prediction problem. We have designed within M-STAP three important

modules, *i.e.,* Two-dimensional Spatial Attention Mechanism (TDSAM), Crowd Flow Spatial Attention Mechanism (CFSAM) and Temporal Attention Mechanism (TAM). In TDSAM, we utilize the multi-head self-attention mechanism to capture the global spatial features of anomaly events and crowd flows of each part of the city in parallel. The extracted citywide feature map is then concatenated with the local spatial feature map extracted by Convolution Operation. We consider the impacts of crowd flow data from different regions on the anomaly events in each time step by CFSAM, and weight the influences of history anomaly events in different time steps for the prediction in TAM. We have evaluated our proposed method with the crowd flows and anomaly events in NYC, Melbourne, and Chicago. The experimental results show that our proposed method works well in multi-region anomaly event prediction problem, outperforming other baselines. Our accurate and robust prediction algorithm can be integrated with other urban computing systems [9, 17] to enhance the preparedness of urban safety management, traffic coordination, and emergency planning.

ACKNOWLEDGMENTS

This project is supported, in part, by the University of Connecticut (UConn) Research Excellence Program (REP) Award (FY20–21 REP).

REFERENCES

[1] Jie Bao, Yu Zheng, and Mohamed F Mokbel. 2012. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proc. ACM SIGSPATIAL*. 199–208.
[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. 2019. Attention augmented convolutional networks. In *Proc. IEEE CVPR*. 3286–3295.
[3] Hancheng Cao, Zhilong Chen, Fengli Xu, Yong Li, and Vassilis Kostakos. 2018. Revisitation in urban space vs. online: A comparison across POIs, websites, and smartphone apps. *Proc. ACM IMWUT* 2, 4 (2018), 1–24.
[4] Zhilong Chen, Hancheng Cao, Huangdong Wang, Fengli Xu, Vassilis Kostakos, and Yong Li. 2020. Will you come back/check-in again? understanding characteristics leading to urban revisitation and re-check-in. *Proc. ACM IMWUT* 4, 3 (2020), 1–27.
[5] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2019. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584* (2019).
[6] Krittika D'Silva, Kasthuri Jayarajah, Anastasios Noulas, Cecilia Mascolo, and Archan Misra. 2018. The role of urban mobility in retail business survival. *Proc. ACM IMWUT* 2, 3 (2018), 1–22.
[7] Yali Fan, Zhen Tu, Yong Li, Xiang Chen, Hui Gao, Lin Zhang, Li Su, and Depeng Jin. 2019. Personalized Context-aware Collaborative Online Activity Prediction. *Proc. ACM IMWUT* 3, 4 (2019), 1–28.
[8] Zipei Fan, Xuan Song, Renhe Jiang, Quanjun Chen, and Ryosuke Shibasaki. 2019. Decentralized Attention-based Personalized Human Mobility Prediction. *Proc. ACM IMWUT* 3, 4 (2019), 1–26.
[9] Zhihan Fang, Yu Yang, Shuai Wang, Boyang Fu, Zixing Song, Fan Zhang, and Desheng Zhang. 2019. MAC: Measuring the impacts of anomalies on travel time of multiple transportation systems. *Proc. ACM IMWUT* 3, 2 (2019), 1–24.
[10] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. 2020. PMF: A privacy-preserving human mobility prediction framework via federated learning. *Proc. ACM IMWUT* 4, 1 (2020), 1–21.
[11] Roger Frigola. 2015. *Bayesian time series learning with Gaussian processes.* Ph.D. Dissertation. University of Cambridge.
[12] Suining He and Kang G Shin. 2019. Spatio-temporal capsule-based reinforcement learning for mobility-on-demand network coordination. In *Proc. WWW*. 2806–2813.
[13] Suining He and Kang G. Shin. 2020. Towards Fine-Grained Flow Forecasting: A Graph Attention Approach for Bike Sharing Systems. In *Proc. WWW*. 88–98.
[14] Chao Huang, Xian Wu, and Dong Wang. 2016. Crowdsourcing-based urban anomaly prediction system for smart cities. In *Proc. ACM CIKM*. 1969–1972.
[15] Chao Huang, Chuxu Zhang, Peng Dai, and Liefeng Bo. 2020. Cross-Interaction Hierarchical Attention Networks for Urban Anomaly Prediction. In *Proc. IJCAI*. 4359–4365.
[16] Renhe Jiang, Xuan Song, Zipei Fan, Tianqi Xia, Quanjun Chen, Qi Chen, and Ryosuke Shibasaki. 2018. Deep ROI-based modeling for urban human mobility prediction. *Proc. ACM IMWUT* 2, 1 (2018), 1–29.
[17] 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 *Proc. ACM SIGKDD*. 2114–2122.

[18] Kenneth Joseph, Chun How Tan, and Kathleen M. Carley. 2012. Beyond "Local", "Categories" and "Friends": Clustering Foursquare Users with Latent "Topics". In *Proc. ACM UbiComp*. 919–926.
[19] Xiangjie Kong, Ximeng Song, Feng Xia, Haochen Guo, Jinzhong Wang, and Amr Tolba. 2018. LoTAD: Long-term traffic anomaly detection based on crowdsourced bus trajectory data. *World Wide Web* 21, 3 (2018), 825–847.
[20] Kundan Krishna, Deepali Jain, Sanket V Mehta, and Sunav Choudhary. 2018. An lstm based system for prediction of human activities with durations. *Proc. ACM IMWUT* 1, 4 (2018), 1–31.
[21] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. 2015. Traffic prediction in a bike-sharing system. In *Proc. ACM SIGSPATIAL*. 1–10.
[22] Xuelian Long, Lei Jin, and James Joshi. 2012. Exploring Trajectory-Driven Local Geographic Topics in Foursquare. In *Proc. ACM UbiComp*. 927–934.
[23] Tatiana Pontes, Marisa Vasconcelos, Jussara Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. 2012. We Know Where You Live: Privacy Characterization of Foursquare Behavior. In *Proc. ACM UbiComp*. 898–905.
[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] Sijie Ruan, Jie Bao, Yuxuan Liang, Ruiyuan Li, Tianfu He, Chuishi Meng, Yanhua Li, Yingcai Wu, and Yu Zheng. 2020. Dynamic Public Resource Allocation Based on Human Mobility Prediction. *Proc. ACM IMWUT* 4, 1 (2020), 1–22.
[26] Shun-Yao Shih, Fan-Keng Sun, and Hung-yi Lee. 2019. Temporal pattern attention for multivariate time series forecasting. *Machine Learning* 108, 8 (2019), 1421–1441.
[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*. 5998–6008.
[28] Huandong Wang, Yong Li, Sihan Zeng, Gang Wang, Pengyu Zhang, Pan Hui, and Depeng Jin. 2019. Modeling spatio-temporal app usage for a large user population. *Proc. ACM IMWUT* 3, 1 (2019), 1–23.
[29] Yilun Wang, Yu Zheng, and Yexiang Xue. 2014. Travel time estimation of a path using sparse trajectories. In *Proc. ACM SIGKDD*. 25–34.
[30] Apichon Witayangkurn, Teerayut Horanont, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2013. Anomalous event detection on large-scale GPS data from mobile phones using hidden Markov model and cloud platform. In *Proc. ACM UbiComp Adjunct*. 1219–1228.
[31] Xian Wu, Yuxiao Dong, Chao Huang, Jian Xu, Dong Wang, and Nitesh V Chawla. 2017. UAPD: Predicting urban anomalies from spatial-temporal data. In *Proc. ECML PKDD*. 622–638.
[32] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 2020. Connecting the Dots: Multivariate time series forecasting with graph neural networks. In *Proc. ACM SIGKDD*. 753–763.
[33] Su Yang, Minjie Wang, Wenshan Wang, Yi Sun, Jun Gao, Weishan Zhang, and Jiulong Zhang. 2017. Predicting commercial activeness over urban big data. *Proc. ACM IMWUT* 1, 3 (2017), 1–20.
[34] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proc. AAAI*, Vol. 32.
[35] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proc. ACM SIGKDD*. 984–992.
[36] Huichu Zhang, Yu Zheng, and Yong Yu. 2018. Detecting urban anomalies using multiple spatio-temporal data sources. *Proc. ACM IMWUT* 2, 1 (2018), 1–18.
[37] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proc. AAAI*. AAAI Press, 1655–1661.
[38] Mingyang Zhang, Tong Li, Hongzhi Shi, Yong Li, Pan Hui, et al. 2019. A decomposition approach for urban anomaly detection across spatiotemporal data. In *Proc. IJCAI*. 6043–6049.
[39] Xiangyu Zhao and Jiliang Tang. 2017. Modeling temporal-spatial correlations for crime prediction. In *Proc. ACM CIKM*. 497–506.
[40] Yu Zheng, Xing Xie, Wei-Ying Ma, et al. 2010. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* 33, 2 (2010), 32–39.
[41] Yu Zheng, Huichu Zhang, and Yong Yu. 2015. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In *Proc. ACM SIGSPATIAL*. 1–10.