Article

An Integrated Graph Model for Spatial–Temporal Urban Crime Prediction Based on Attention Mechanism

Miaomiao Hou [1], Xiaofeng Hu [1,*], Jitao Cai [2], Xinge Han [2] and Shuaiqi Yuan [3]

1   School of Information Technology and Cyber Security, People's Public Security University of China, Beijing 100038, China; 201621220042@stu.ppsuc.edu.cn
2   School of Emergency Management and Safety Engineering, China University of Mining and Technology, Beijing 100083, China; sqt1910101006@student.cumtb.edu.cn (J.C.); bqt2110104039@student.cumtb.edu.cn (X.H.)
3   Safety and Security Science Section, Faculty of Technology, Policy and Management, TU Delft, 2628 BX Delft, The Netherlands; s.yuan-2@tudelft.nl
*   Correspondence: huxiaofeng@ppsuc.edu.cn

Abstract: Crime issues have been attracting widespread attention from citizens and managers of cities due to their unexpected and massive consequences. As an effective technique to prevent and control urban crimes, the data-driven spatial–temporal crime prediction can provide reasonable estimations associated with the crime hotspot. It thus contributes to the decision making of relevant departments under limited resources, as well as promotes civilized urban development. However, the deficient performance in the aspect of the daily spatial–temporal crime prediction at the urban-district-scale needs to be further resolved, which serves as a critical role in police resource allocation. In order to establish a practical and effective daily crime prediction framework at an urban police-district-scale, an "online" integrated graph model is proposed. A residual neural network (ResNet), graph convolutional network (GCN), and long short-term memory (LSTM) are integrated with an attention mechanism in the proposed model to extract and fuse the spatial–temporal features, topological graphs, and external features. Then, the "online" integrated graph model is validated by daily theft and assault data within 22 police districts in the city of Chicago, US from 1 January 2015 to 7 January 2020. Additionally, several widely used baseline models, including autoregressive integrated moving average (ARIMA), ridge regression, support vector regression (SVR), random forest, extreme gradient boosting (XGBoost), LSTM, convolutional neural network (CNN), and Conv-LSTM models, are compared with the proposed model from a quantitative point of view by using the same dataset. The results show that the predicted spatial–temporal patterns by the proposed model are close to the observations. Moreover, the integrated graph model performs more accurately since it has lower average values of the mean absolute error (MAE) and root mean square error (RMSE) than the other eight models. Therefore, the proposed model has great potential in supporting the decision making for the police in the fields of patrolling and investigation, as well as resource allocation.

Keywords: urban crime; graph convolutional network; attention mechanism; spatial–temporal prediction; LSTM network

1. Introduction

In recent years, spatial–temporal crime prediction technology has been rapidly developed. With deriving data that include the crime incidents number, population density, weather variables, etc., spatial–temporal patterns of assault, robbery, theft, or other types of crimes can be predicted with the help of machine learning (especially deep learning) and other methods [1–8]. It provides references and predictions about when and where the crime hotspot would be to the police in advance, so it contributes to crime prevention as well as better police resources allocations. With respect to practicality, the accuracy of crime prediction can be regarded as the most important indicator.

In order to establish accurate spatial–temporal crime prediction models, support vector machine (SVM), random forest (RF), and other machine learning algorithms were adopted to predict the spatial and temporal distributions of different types of crimes. For example, Ingilevich and Ivanov [9] made a comparative study about three predictive models (linear regression, logistic regression, and gradient boosting) and pointed out that the accuracy of the gradient boosting model (which is a typical machine learning model) was much higher than that of the other two statistics-based models. Yu et al. [10] proposed a new ST-cokriging method for crime prediction on weekly, biweekly, and quadweekly scales, and the validation results showed that the minimum root mean square error (RMSE = 0.145) was on the weekly scale. Compared with spatial–temporal prediction, the time series prediction of crime has been studied by more scholars by using machine learning frameworks. Dash et al. [11] utilized various public city data to predict the time series of the crime incidents number of Chicago in the US, and reported that the prediction performance of the support vector regression was higher than that of the polynomial and auto-regressive methods. Chen et al. [12] applied a support vector machine method based on fuzzy information granulation to analyze and forecast the crime rates in a city of China, and the results indicated that the accuracy of this model is much better than that of ARIMA. Pillai [13] adopted machine learning models extreme gradient boosting (XGBoost), random forest, etc., to predict the violation and narcotics crimes in Chicago, and found that XGBoost outperformed all of the other models, with the highest R^2 value of 88% and RMSE value of 2.57 towards crime incidents. In general, machine-learning-based methods are more successful in time series predictions than spatial–temporal predictions of crimes. One of the most important shortcomings may be that the ability of the spatial feature extraction of these models is inadequate.

In the last decades, deep neural networks became a fruitful approach to address the problems of complex feature engineering with their powerful self-learning capacity. Deep neural networks have been stimulated and widely applied in the fields of computer vision [14,15]. With deep learning, complicated spatial–temporal features of crimes can be captured and utilized to predict the future patterns of crimes. Li et al. [16] forecasted the yearly property crime incidents number of a city (with a spatial resolution of 100 m × 100 m) that is located in the south of China by using back propagation (BP) neural networks and genetic algorithms. Zhang et al. [17] studied the hourly prediction of crime patterns in the city of Suzhou in China based on the spatial–temporal residual networks (ST-ResNet) model: an adaptive spatial resolution method was proposed to find out the best spatial resolution for crime risk prediction, and the results pointed out that a 2.4 km spatial resolution can obtain the best performance for crime prediction (RMSE = 7.81). Qian et al. [18] established the GeST model based on grid division to predict the number of theft crimes in New York City from 2011 to 2018: the experimental result on the test set was better than ARIMA, LSTM, and other models. Zhang et al. [19] designed a new optimization modeling method for crime prediction based on the grid management concept by using a BP neural network algorithm, which was validated by the data of four different crimes in Chicago, and the results showed that the RMSE of this method fell by approximately 9% on average compared with the traditional meshless models. Han et al. [20] proposed an integrated model of crime prediction by combining LSTM with ST-GCN to forecast the spatial–temporal crime risk of urban communities in Chicago, and the result showed that this model had been verified by practical examples and achieved a superior prediction to Ridge, random forest, and LSTM models.

As reported by the previous literatures, deep learning methods seem to be more suitable for spatial–temporal crime prediction than the traditional machine learning algorithms. However, for the purposes of supporting decision making for the police, there are still three main gaps. Firstly, most of the previous studies focused on yearly, monthly, or weekly scales rather than a daily scale due to the insufficient accuracy of the models. A few studies focused on a daily (even hourly) scale, but usually obtained relatively larger errors in terms of RMSE or MAPE [9,10,16–20]. With respect to security inspection and patrolling, as well

as other associated jobs, police may be more interested in the effective daily crime prediction results. Secondly, a considerable number of studies are based on regular spatial–temporal grids for spatial division [9,16–19], but this is inconsistent with actual administrative areas. Generally, policing strategies should be consistent within the same administrative area (such as the community or police district). As a risk compensation element, it may affect the spatial–temporal distribution of crimes. Thus, the grid-based deep learning frameworks may have limited accuracy in terms of spatial–temporal crime predictions. Thirdly, spatial–temporal crime prediction models should also pay attention to convenience and feasibility. However, most of the module connections of the present models are "offline", which means that the spatial prediction module and the temporal prediction module are separated, not "online". This may affect its convenience in practical applications.

To address these shortcomings, an "online" integrated graph model based on the attention mechanism for spatial–temporal crime prediction is proposed in this study. This model combines a residual neural network (ResNet), graph convolutional network (GCN), and LSTM to extract spatial–temporal features from the crime data. Moreover, the topological relationship among different urban districts is also captured, and the attention mechanism is incorporated into this model. The attention probability distribution can give high weight to the critical information while ruling out irrelevant information in the proposed model. Finally, feature fusion is completed through the weighted calculation of multiple features; meanwhile, the "offline" problem is solved as well. To validate this "online" integrated graph model, the daily theft and assault data in a police-district scale of the city of Chicago in the US from 1 January 2015, to 7 January 2020 are used. Indices of mean absolute error (MAE) and RMSE are employed to quantitatively evaluate this model. Furthermore, comparative analysis is also carried out in this study with other eight models applied in previous studies.

The rest of this paper is organized as follows. Section 2 describes the investigated area; presents the data used for the model validation; elaborates on the spatial–temporal crime prediction algorithm and the validation schemes. Section 3 is about the prediction results and discussions. Finally, Section 4 summarizes our main contributions and puts forward a set of recommendations for further study.

2. Materials and Methods

2.1. Study Area and Data Description

The city of Chicago was selected as the investigated area. It is the most densely populated city in Illinois, and the third-largest city in the United States, following New York City and Los Angeles. By the end of 2020, the estimated population of Chicago is 2,746,388 (Source: https://www.census.gov/, accessed on 1 October 2020). Chicago is in the center of the North American continent, and it is one of the most important international financial centers. As a typical large city, Chicago suffers from a large number of crimes, including the types of assaults, burglary, theft, etc., so it is very suitable for validating crime prediction models. A number of previous studies also chose this city to validate their proposed crime prediction methods, which can provide comparisons for this research.

This city has 22 police districts that overlap with 77 communities (as shown in Figure 1). In some of the police districts, one police station may serve multiple communities, while it is also possible that one community is covered by more than one police district. As the smallest cell of a city, the community is usually selected as the spatial unit to predict the spatial–temporal patterns of crime in previous studies [6,20–23]. However, in this study, spatial areas were divided by police districts. As is well known, crime risk is not only affected by population density, traffic environment, economic level, etc., but is also considerably influenced by the work of the police (such as patrolling, security check, etc.), which should be considered as a non-ignorable risk compensation factor. Within the same police district, the police strategy usually has spatial homogeneity. By contrast, the strategies may be varied between different police districts so that the impacts of police on crime risks may be quite different. This is why we chose police districts (rather than

communities) to spatially divide the city. Another reason is that the spatial–temporal crime prediction results in a specific police district that can be applied directly to supporting the decision making with respect to police resource allocation within this district. As shown in Figure 1, the 22 police districts of Chicago are numbered from 1 to 25, excluding 13, 21, and 23, in which, the largest and the smallest police districts have areas of 816.6 km^2 and 79.9 km^2, respectively.
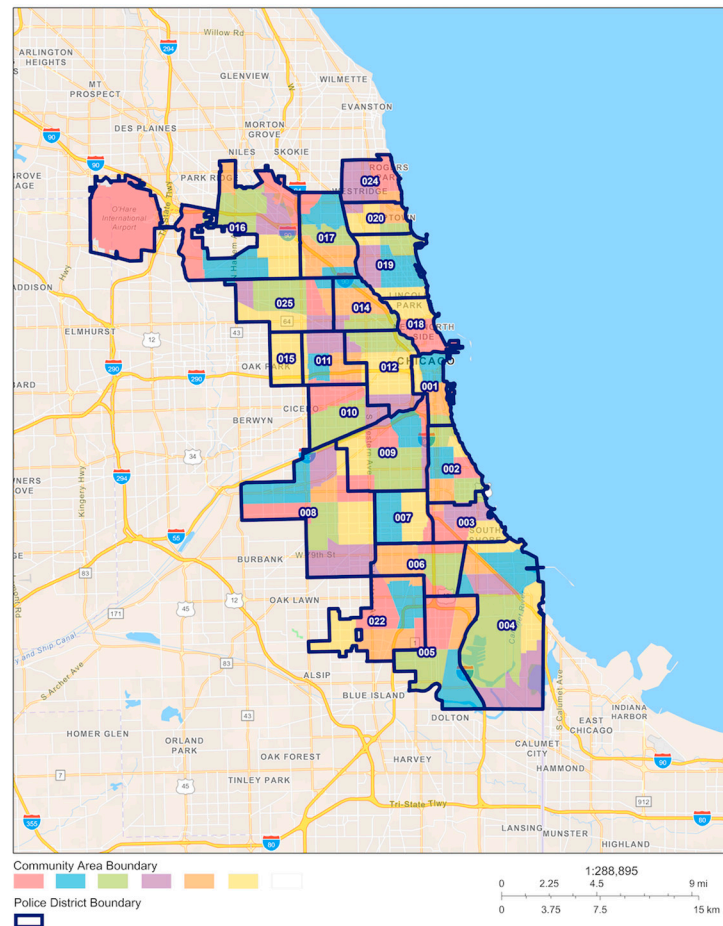

Figure 1. Map of the police districts in Chicago.

The dataset used for training and testing the established model was derived from the Chicago open data portal (https://data.cityofchicago.org/, accessed on 1 October 2020), which is a publicly available data search and exploration platform developed (and currently managed) by the University of Chicago's Urban Center for Computation and Data. The crime data were collected from the "Crimes-2001 to present" dataset, a real-life collection of instances describing crime events in Chicago from 2001 to the present in which the repository is updated every week. From the "Crimes-2001 to present" dataset, we collected theft (as one of the typical property crimes) and assault (as one of the typical violent crimes) incidents numbers within the 22 police districts over five years (1833 days) from 1 January 2015, to 7 January 2020. In the data pre-processing stage, missing data were dropped, and the number of theft and assault incidents in the 22 police districts of Chicago was counted at a daily scale. A total of 404,269 pieces of valid data were finally obtained, including 308,020 pieces of theft data and 96,249 pieces of assault data. Figure 2 shows the temporal distribution of theft and assault incidents numbers in the 22 police districts of Chicago. As shown in Figure 2, both seasonality and daily variations of theft and assault can be observed.
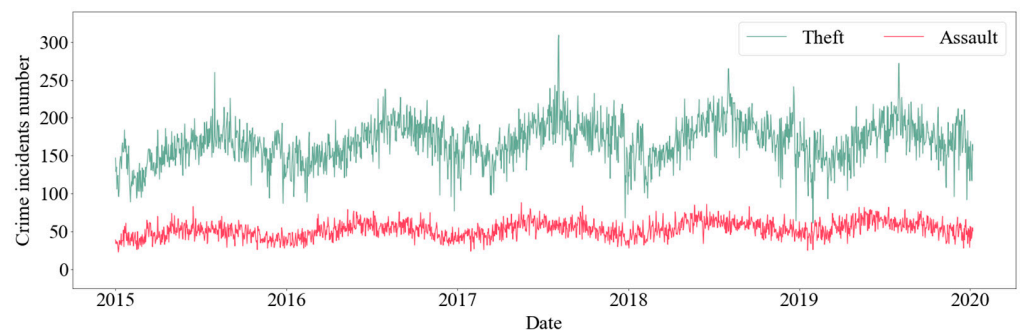
Figure 2. Daily number of theft and assault incidents in the 22 police districts of Chicago from 1 January 2015, to 7 January 2020. Blue line and green line represent theft and assault incidents numbers, respectively.

Figure 3 shows the spatial patterns of theft and assault incidents numbers in Chicago. The theft incidents in Chicago are mainly concentrated in the northeast areas, whereas assault incidents are indicated by hot areas in the southeast. According to this figure, the accumulated theft number in the 18th police district reaches approximately 30000, which is much larger than that in other districts, indicating that theft has a significant spatial aggregation. The descriptive statistics of theft and assault incidents numbers in the 22 police districts are listed in Table A1. The number of theft incidents (M = 18.754, SD = 7.397) is larger than that of assault incidents (M = 1.939, SD = 1.472). Here, M is the crime incidents number's mean value, and SD indicates the crime incidents number's standard deviation.

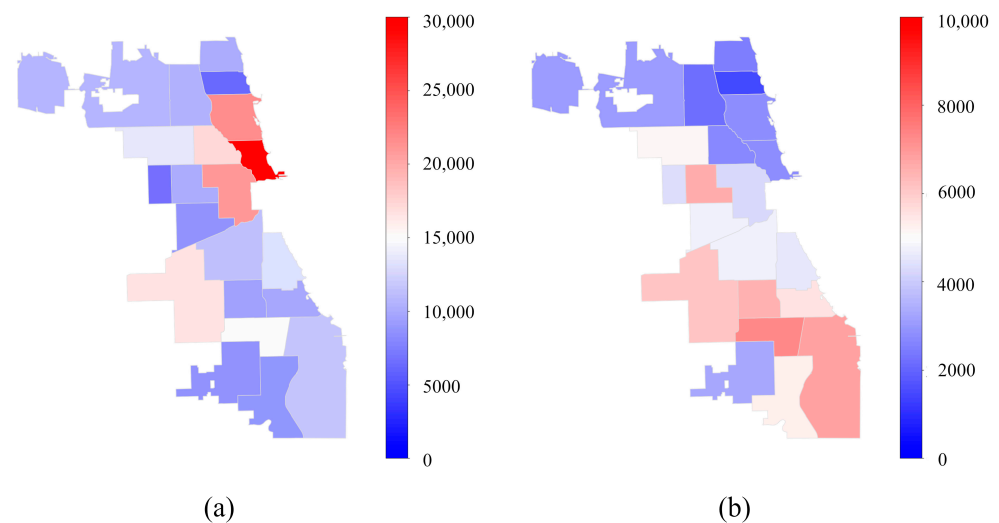

(a)   (b)

Figure 3. The spatial patterns of theft and assault incidents numbers in the 22 police districts of Chicago from 1 January 2015, to 7 January 2020. Panel (**a**) shows the distribution of theft incidents numbers, whereas panel (**b**) is the spatial pattern of assault.

Topological graph of the 22 police districts in Chicago was also used to train the model. According to the neighborhood relationship among police districts of Chicago, the topological map of adjacent police districts is obtained as shown in Figure 4.

Previous studies demonstrated that, besides the spatial–temporal features of crimes, weather variables are also used as external features to build the deep learning crime prediction models [20,22,24–26]. The reason is that the temperature and relative humidity proved to have relationships with some types of crimes, including assault, burglary, robbery, rape, and so on [27–29]. Heat stress indices can well capture the combination of the impacts of temperature and relative humidity on crime rates. Therefore, we chose the heat stress discomfort index (DI) [20,27,28] as one of the external features, and the description of DI is

presented in Appendix B. Apart from DI, the features "Weekend" and "Holiday" were also applied as the external features. The descriptions of these features are listed in Table 1.
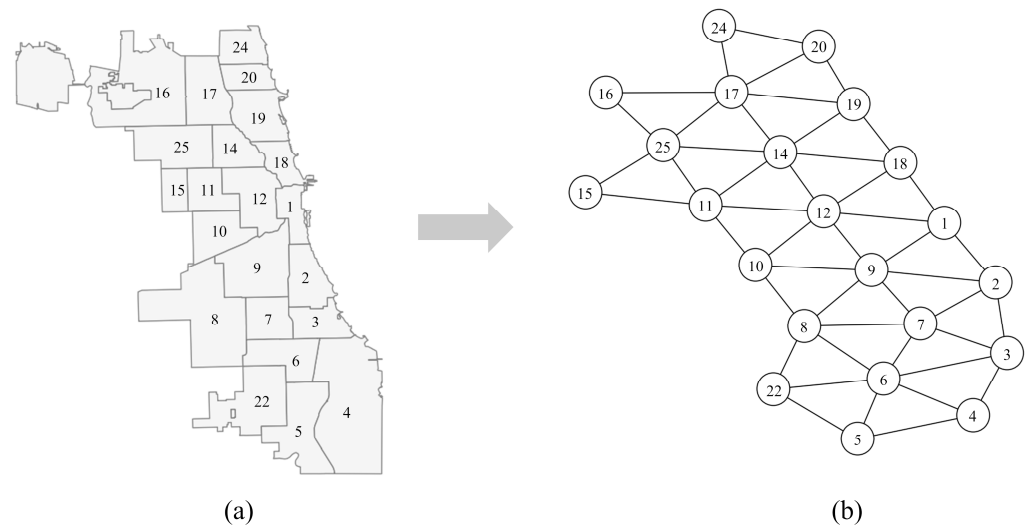

(a)   (b)

Figure 4. The administrative map and topological map of the 22 police districts in Chicago. Panel (**a**) shows the administrative map, whereas panel (**b**) is the topological map.

Table 1. The brief descriptions of external features.

Feature Name	Description
Weekend	The feature "Weekend" represents whether the day is a weekday or a weekend.
Holiday	The feature "Holiday" represents whether the day is a holiday or not.
DI	The heat stress index DI is calculated according to temperature, humidity, and wind speed (as shown in Appendix B).

2.2. Model Framework

In this study, an integrated graph model based on attention mechanism was developed for spatial–temporal crime prediction that can extract the spatial–temporal features of crime incidents numbers and combine them with the external features. The architecture of the proposed model is presented in Figure 5. The model consisted of two modules: (i) the spatial–temporal features extraction module and (ii) the feature fusion and training module. In the feature extraction module, the spatial–temporal features of crimes, the topological features of the police districts, and the external features, such as "DI", "Holiday", etc., were extracted from the data. In the feature fusion and training module, all of the features were integrated into the LSTM attention to constitute a well-trained model. Finally, the test data set was employed to evaluate the prediction performance of the model, and the results were compared with those of other algorithms by using the same test set.
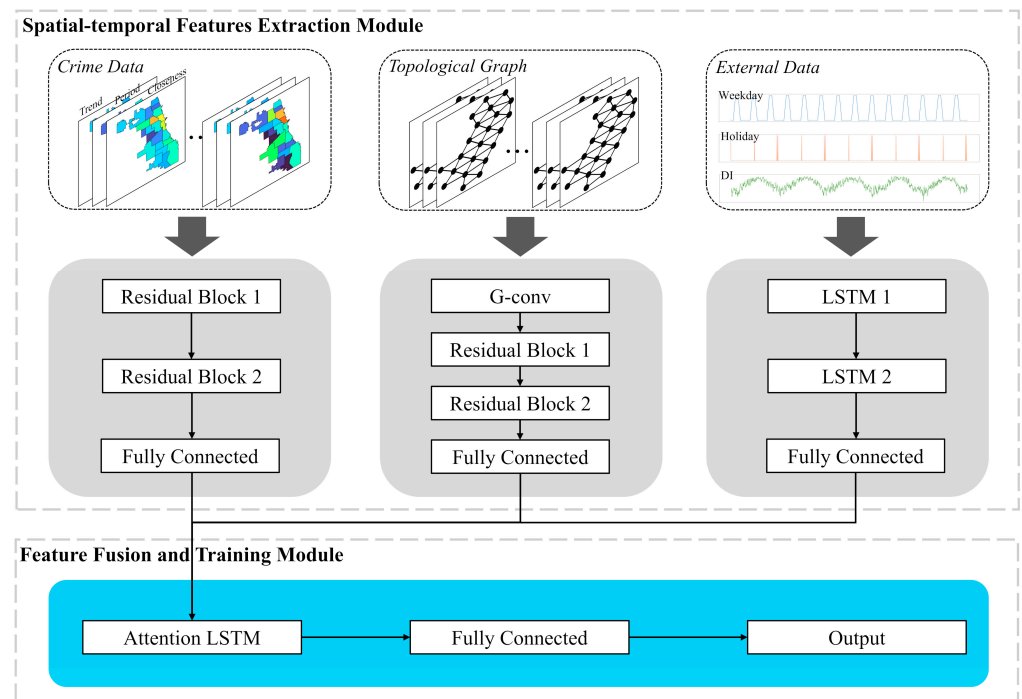
Figure 5. The architecture of the proposed integrated graph model.

2.2.1. Spatial–Temporal Features Extraction Module

To establish the spatial–temporal features extraction module, deep residual neural networks (ResNet) were employed. ResNet was first proposed in 2015 by He et al. [30] to solve gradient disappearance, gradient explosion, network degradation, and other tricky problems induced by increasing the number of layers in convolutional neural network (CNN). ResNet is a stack of "residual blocks" as shown in Figure 6. As shown in panel (a), x represents the residual block input, F(x) + x represents the output, weight layer is a convolution layer, and rectified linear unit (ReLU) represents an activation layer [31].
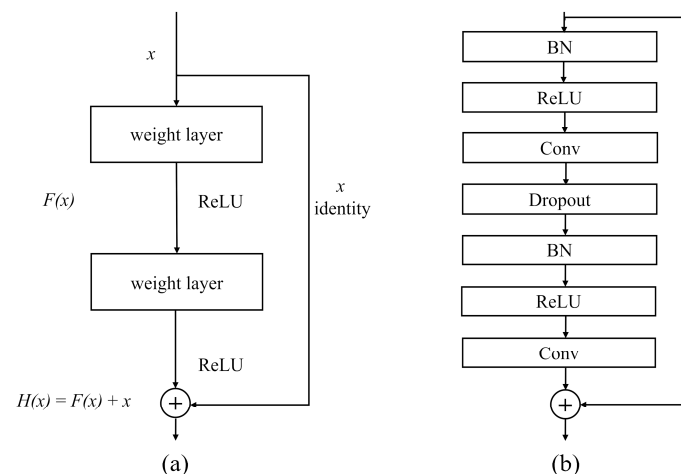
Figure 6. Two types of residual blocks. Panel (**a**) shows the original residual block, whereas panel (**b**) is the improved residual blocks.

In this study, an improved residual block was adopted to enhance the training speed and also the accuracy, which is illustrated in the right panel of Figure 6 [32]. The improved residual block adjusts the batch normalization (BN) and ReLU to the front of the convolutional layer (Conv). BN is used for data standardization, which can significantly accelerate the convergence of network training. By inputting the processed data into ReLU

for activation, the sparsity of the network and the nonlinear relationship between layers can be enhanced; meanwhile, the impact of the overfitting problem can be alleviated. The activated data were converted to the Conv layer for feature extraction. By adding a dropout layer between Conv layers, the probability of network overfitting can be reduced. Moreover, three major features were designed to capture the spatial–temporal properties from the crime data, including closeness features, periodic features, and trend features [17,20,26,33]. Specifically, to predict the crime pattern on a certain day, the number of daily crime incidents three days before was extracted as the closeness features. The number of daily crime incidents on the 7th day, the 14th day, and the 21st day before the day was extracted as the periodic features, and the number of daily crime incidents on the corresponding dates in the past three years was extracted as the trend features. The abovementioned crime data were input into two improved residual units, and a full connection layer was used to connect with the next module.

Crime incidents in a specific district are not only temporally associated with the number of crime incidents in the past but also spatially impacted by its surrounding districts. In this study, graph convolutional network was applied to capture the network topology dependency and describe the spatial correlation of crimes more accurately [34]. The forward propagation of GCN is calculated as follows:

$$f(\boldsymbol{X},\boldsymbol{A}) = \mathrm{ReLU}(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W} + \boldsymbol{b}) \tag{1}$$

where $\boldsymbol{X}$ is the feature matrix, $\boldsymbol{A}$ is the adjacency matrix of the graph, $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$ is the adjacency matrix with self-circulation added, $\boldsymbol{D}$ is the degree matrix, and $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weight matrix and the bias, respectively [35–37]. In this paper, ResNet was selected as the carrier of GCN to build the network, and the topological features were input into the GCN layer and two improved residual units. Moreover, a full connection layer was used to connect the feature fusion module. As for external features, the processed data were input into two LSTM layers, and a full connection layer was also used to transmit the results into the next module.

2.2.2. Feature Fusion and Training Module

Compared with the traditional recurrent neural network (RNN) model, LSTM adds memory units to each hidden layer neural unit to obtain the controllable memory information on the time series [38–40]. However, when coping with enormous multi-dimensional and multivariable data, the model of LSTM may ignore some vital time-series information in practical use, resulting in poor performance. Therefore, we adopted the attention mechanism based on LSTM to fuse the features. This method can break the limitation of the encode–decoder architecture and the fixed-length internal representation. The attention mechanism can be used to simulate the attention mechanism of human brains [41,42]. Attention LSTM retains the intermediate states of the LSTM encoder and selectively learns these intermediate states through the training model. The LSTM model can be combined with the attention mechanism to give different weights to the input characteristics of the LSTM and highlight the critical influencing features without increasing the calculation and storage overhead of the model, helping LSTM to make more accurate judgment [43]. Let matrix $\boldsymbol{X} \in \boldsymbol{R}^{m \times n}$ be the LSTM output, where m is the timesteps, and n represents the number of features in each timestep. Then, the mathematical formulation of the attention-based output ($O_{\mathbf{Attention}}$) is presented as follows:

$$\boldsymbol{A} = f(\boldsymbol{W} \circ \boldsymbol{X} + \boldsymbol{b}) \tag{2}$$

$$O_{\mathbf{Attention}} = \boldsymbol{A} \circ \boldsymbol{X} \tag{3}$$

where $\boldsymbol{A}$ is a weight matrix with the same shape as $\boldsymbol{X}$, "$\circ$" represents the Hadamard product, f represents the fully connected layer, and $\boldsymbol{W}$ and $\boldsymbol{b}$ denote the weight matrix and the bias, respectively [44]. In this module, feature fusion was completed through the weighted

calculation of multiple features. The spatial prediction module and the temporal prediction module were not separated in the proposed model. Thus, the "offline" problem was solved as well. The results from the spatial–temporal features extraction module were input into the attention LSTM, which was followed by another fully connected layer to output the final results.

2.3. Case Study

2.3.1. Model Configuration

The proposed model was implemented by Keras and TensorFlow. The system hardware of the server we used included a CPU with two cores (Intel Xeon E5 Core *20), four GPUs (Nvidia Tesla P100 16GB), four memory modules (Kingston 64GB 2666MHz), and three hard disks (4TB). Both theft and assault incidents data from 1 January 2015 to 31 December 2019 were used to train this model, and the incident data from 1 January to 7 January 2020 were used as the test set. The validation split rate was set as 0.2 to calibrate the model. The dimensions of tensors flowing in the proposed model are shown in Figure 7. In terms of the crime data processing part, the first residual block in the spatial–temporal features extraction module had 32 filters, whereas the second one had 64 filters. The kernel size was 3×3 and the fully connected layer consisted of 22 neurons. The topological features processing part passed through a GCN layer first, and then the remaining configuration was the same as the crime data processing part. Two residual blocks had 32 and 64 filters, respectively. The kernel size was 3×3 and the fully connected layer consisted of 22 neurons in that case. As for the external data processing part, the two LSTM layers consisted of 128 and 276 neurons, respectively, and the fully connected layers comprised 22 neurons. For feature fusion and training module, the attention LSTM and final fully connected layers consisted of 128 and 22 neurons, respectively. Furthermore, we adopted end-to-end training to optimize the model. The mean squared error (MSE) was used as the loss function. The optimizer was "NAdam", with a learning rate of 0.001.
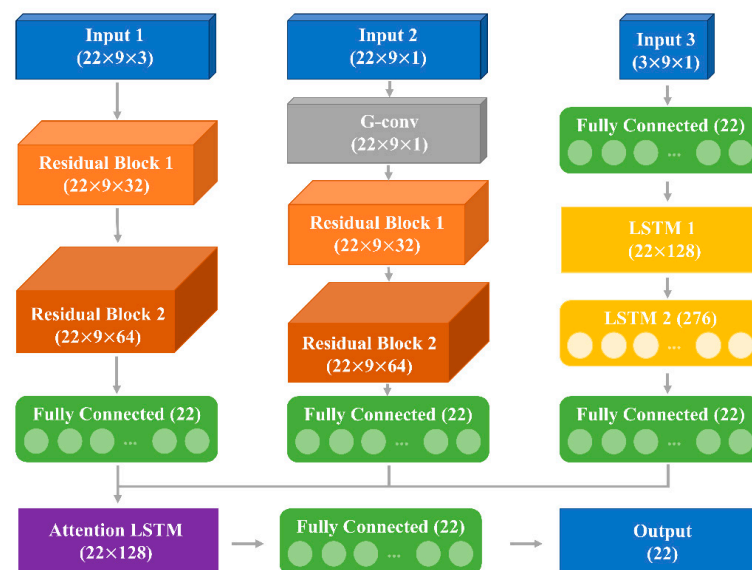

Figure 7. Dimensions of tensors flowing in the proposed model.

2.3.2. Baseline Models

In this study, eight algorithms were compared with our model to verify its performance, including ARMIA [45], ridge regression [46], support vector regression (SVR) [47], random forest [48], XGBoost [49], LSTM [50], CNN [51], and Conv-LSTM [52]. The specific configurations are shown in Table 2.

Table 2. The specific configurations of baseline models.

Baseline Models.	Specific Configurations
ARIMA	The best ARIMA results are obtained automatically by Expert Modeler in the Statistical Package for the Social Sciences (SPSS) software.
Ridge Regression	The hyperparameters of ridge regression are obtained automatically by the RidgeCV class in the scikit-learn library.
SVR	The kernel of SVR is set as a radial-basis function, the regularization parameter C is set as 1.0, and the tolerance for stopping criterion is set as 0.001.
Random Forest	The number of trees in random forest is set as 100, and the maximum tree depth is set as 10.
XGBoost	The number of iterations is set as 100, and the maximum tree depth is set as 10.
LSTM	The LSTM has two kernel layers (each containing 100 neurons), and the learning rate is set as 0.0001.
CNN	The CNN has two kernel layers, the number of filters is set as 32 and 64, the size of the kernel is set as 3×3, the input length is set as 12, and the learning rate is set as 0.005.
Conv-LSTM	The number of Conv-LSTM layers is set as 2 with 32 and 64 filters, respectively, the kernel size is 3×3, the input length is set as 12, and the learning rate is set as 0.005.

2.4. Evaluation Metrics

To evaluate the prediction performance of the proposed model, MAE and RMSE were utilized as evaluation metrics. The mathematical formulation of MAE and RMSE is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{5}$$

where Y_i is the observed value, $\hat{Y}_i$ is the predicted value, and *n* is the prediction sample numbers. Smaller value of MAE and RMSE indicates higher prediction accuracy of the model [53,54]. In this study, nine prediction models were trained and tested based on the same data set.

3. Results and Discussion

To analyze the experimental results and evaluate the performance of the proposed model, data in the 22 police districts from 1 January 2020, to 7 January 2020 (7 days) were selected to build the test set. Figure 8 (for theft) and Figure 9 (for assault) show the spatial patterns of the observed crime incidents, prediction of crime incidents numbers, and absolute error (AE) in three representative days (namely 1, 4, and 6 January 2020), as well as the cumulative values from 1 to 7 January 2020. Among the three selected days, 1 January 2020 is a holiday, 4 January 2020 is a weekday, and 6 January 2020 is a weekend.
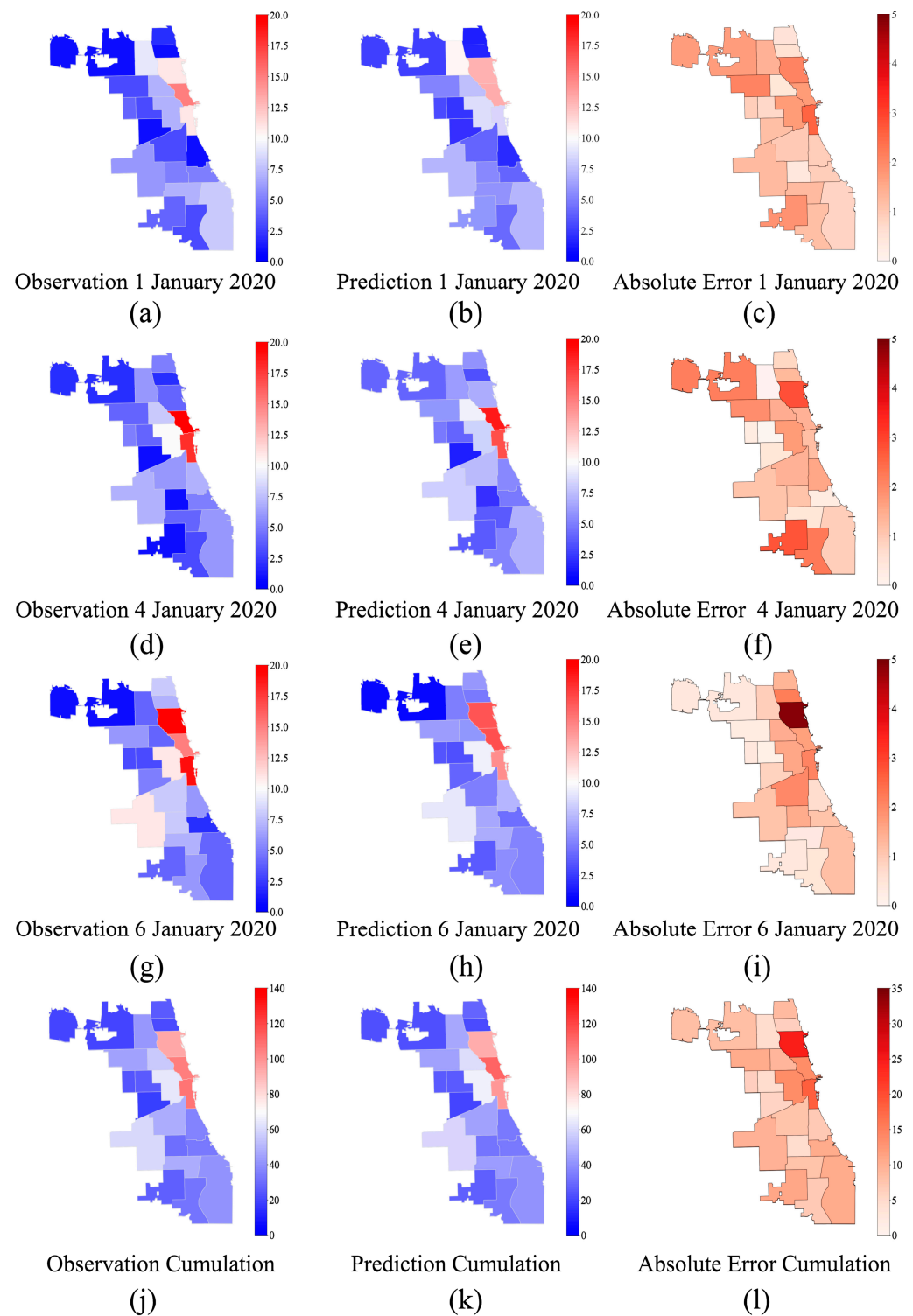
Figure 8. Spatial patterns of the predicted theft incidents numbers of different police districts in Chicago. The left panels (**a,d,g,j**) show the observed crime distributions, the middle panels (**b,e,h,k**) represent the predicted patterns, and the right panels (**c,f,i,l**) are about the AE values between prediction and observation. Panels from top to bottom represent values on January 1, 4, and 6, 2020, and that of the cumulations from 1 to 7 January 2020.

Figure 9. Spatial patterns of the predicted assault incidents numbers of different police districts in Chicago. The left panels (**a,d,g,j**) show the observed crime distributions, the middle panels (**b,e,h,k**) represent the predicted patterns, and the right panels (**c,f,i,l**) are about the AE values between prediction and observation. Panels from top to bottom represent values on January 1, 4, and 6, 2020, and that of the cumulations from 1 to 7 January 2020.
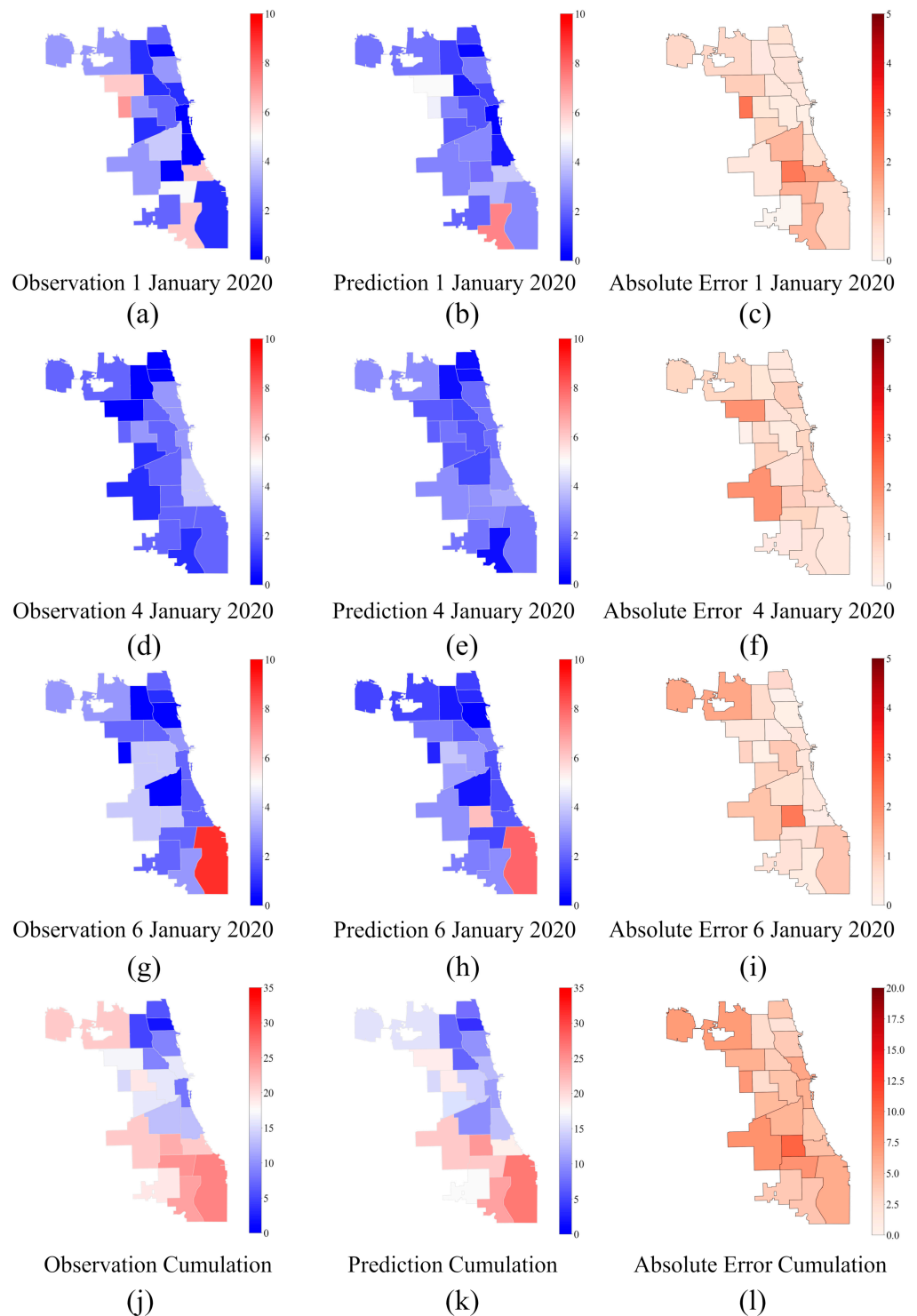
As shown in Figure 8, the predicted patterns of theft are very close to the real distributions, and, as shown in the maps, the proposed model accurately captures the spatial hot areas of theft, not only on the selected three days but also for the cumulated results. Specifically, on 1 January 2020, the 1st district has the max AE (2.57) among the 22 police

districts. As for 4 January 2020 and 6 January 2020, the 19th district witnesses the max AE of 2.80 and 4.87, respectively. Moreover, the 19th district also has the max AE (24.81) for the cumulative values. In terms of the 22 police districts, the average AE on 1 January 2020, 4 January 2020, 6 January 2020, and the cumulations of seven days are 1.24, 1.24, 1.22, and 9.63, which indicates that the proposed model has reliable accuracy in this experiment. The spatial patterns of theft show that the 1st district and the 19th district have larger absolute errors than others. As investigated, both the above two districts are in the northeast of Chicago, which are the main tourist attractions, with many downtown areas of this city. Thus, it is one of the most crowded areas and is usually reported as a hotspot of theft, so various factors, such as the population density, unemployment rate, traffic condition, etc., may comprehensively influence the crime rates, which increase the difficulty in accurately forecasting the daily theft numbers in these districts.

With regard to assault, Figure 9 reports that the predicted spatial–temporal patterns on the test set have close agreements with the observations. Specifically, the spatial hot areas of assault are captured by the proposed prediction model in terms of the cumulative values. Error distributions in the right panels of Figure 9 points out that the 15th district witnesses the max AE (2.31) among the 22 police districts on 1 January 2020, the 8th district has the max AE (1.84) on 4 January 2020, and the 7th district has the max AE of 2.21 on 6 January 2020. The spatial average AE on 1, 4, and 6 January are 0.88, 0.71, and 0.65, proving that this model can accurately forecast the spatial pattern of assault, with an average error lower than one incident. As for the spatial patterns of assault, the police districts with large absolute errors are mainly in the southwest of Chicago, especially the 7th district. The 7th district has a lot of violent incidents due to its complicated ethnic composition and extreme disparity in income. Therefore, the reason why the AE in the 7th district is larger than other places may be that a variety of social factors are associated with crimes, which adds more uncertainty to the prediction of crimes.

Figures 10 and 11 show the prediction of theft and assault incidents numbers of the 22 police districts in Chicago, respectively. As shown in Figures 10 and 11, the predicted value is quite close to the actual value in terms of both theft and assault. Though heavy fluctuations are observed in theft incidents in the 1st, 18th, and 19th districts and in assault incidents in the 4th, 7th, and 22nd districts, this model can also predict the temporal trends with limited errors. Thus, from these two figures, we can find that the temporal patterns of both theft and assault are accurately predicted on the test set, showing a reliable performance of this integrated graph model.

In order to quantitatively examine the overall performance of this model, evaluation metrics (MAE and RMSE) were used. The MAE and RMSE of ARIMA, ridge regression, SVR, random forest, XGBoost, LSTM, CNN, and Conv-LSTM were also calculated as comparisons. In the comparative experiments, the same dataset and the extracted features were used among all of the models. The obtained results are shown in Table 3. The average values of the indices are lower than two incidents for theft (MAE = 1.3778, RMSE =1.6318) and one incident for assault (MAE = 0.7457, RMSE =0.8851), which indicates that this model has a good performance in spatial–temporal crime prediction. Compared with other models, the proposed model has a better performance in terms of both MAE and RMSE since the second lowest MAE (RMSE) of theft is 1.5574 (1.8736) from Conv-LSTM, whereas that of assault is 0.8269 (0.9863), which is also from Conv-LSTM. In addition, the results in this table also point out that the performance of ARIMA, ridge regression, SVR, random forest, or XGBoost is significantly worse than the deep learning models LSTM, CNN, and Conv-LSTM. The reason may be that it is difficult for SVR (and the other statistics-based and traditional machine learning models) to capture the complex features from the spatial–temporal crime data. For the deep learning models, the performance of LSTM is considerably worse than CNN and Conv-LSTM. As is known, the latter two are good at exploring the spatial autocorrelation of the variables and extracting advanced features of the daily spatial distributions of crimes. Since spatial information usually plays an important role in spatial–temporal crime prediction, CNN and Conv-LSTM are better

suited to this scenario than original LSTM, which is only able to examine the temporal patterns. Interestingly, the integrated graph model proposed in this study has a much better performance than the three other deep learning models. The reason may be that the three other deep learning models are based on regular spatial–temporal grids for spatial division. However, many latent spatial correlations among the real administrative districts (such as the policing districts) cannot be well captured, which may adversely affect the prediction performance of such models. By contrast, the proposed model extracts the topological relationships of the 22 police districts in the city using ResNet and GCN. The attention mechanism is also investigated to fuse the complicated features through the weighted calculations. This indicates that the integrated graph model is a practical and effective solution for spatial–temporal crime prediction and it is also convenient and feasible with respect to practical applications.
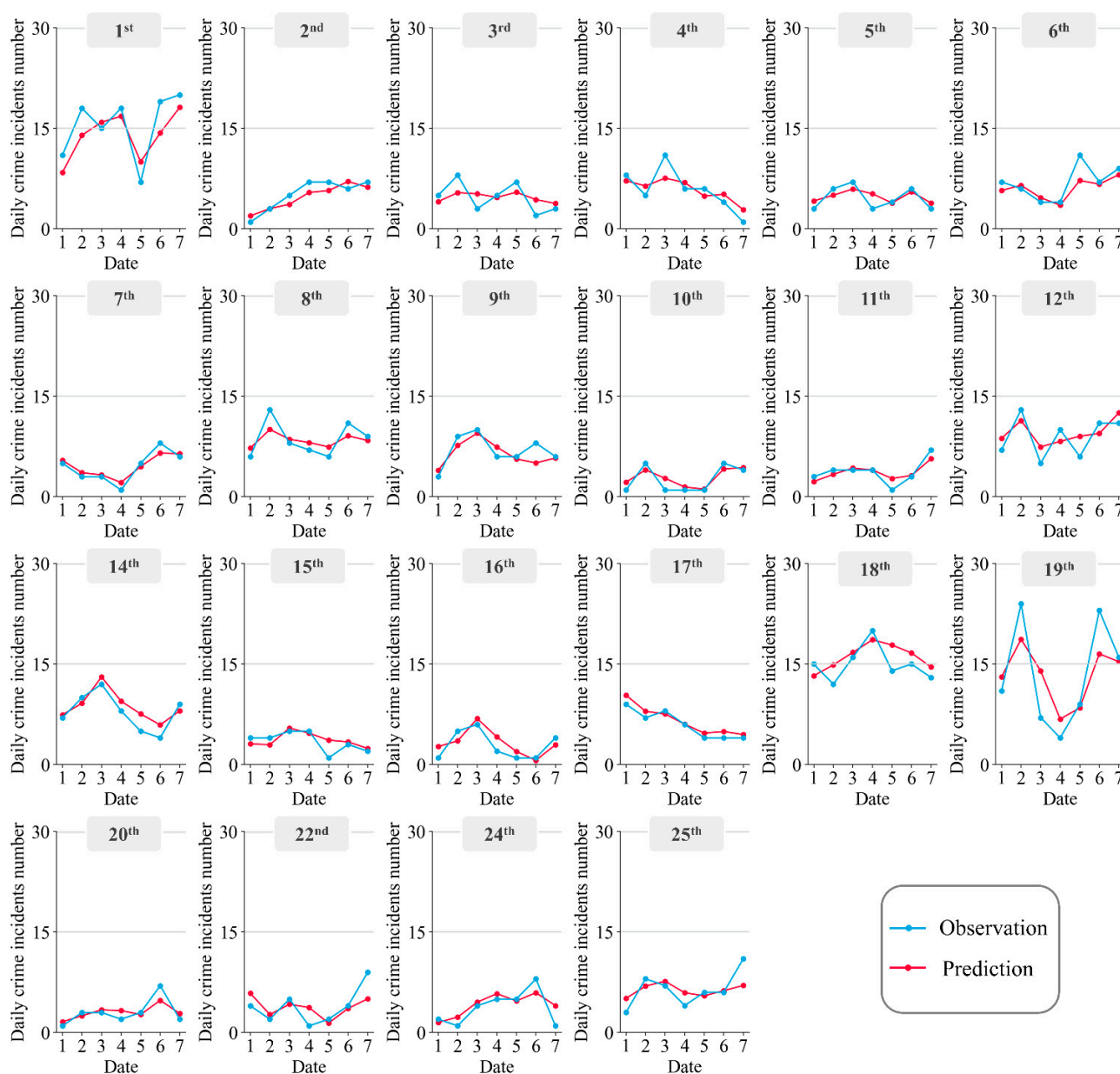

Figure 10. The prediction of daily theft incidents numbers of the 22 police districts in Chicago. Blue lines and points represent the observed daily crime numbers, whereas red lines and points represent the predictions.

Figure 11. The prediction of daily assault incidents numbers of the 22 police districts in Chicago. Blue lines and points represent the observed daily crime numbers, whereas red lines and points represent the predictions.
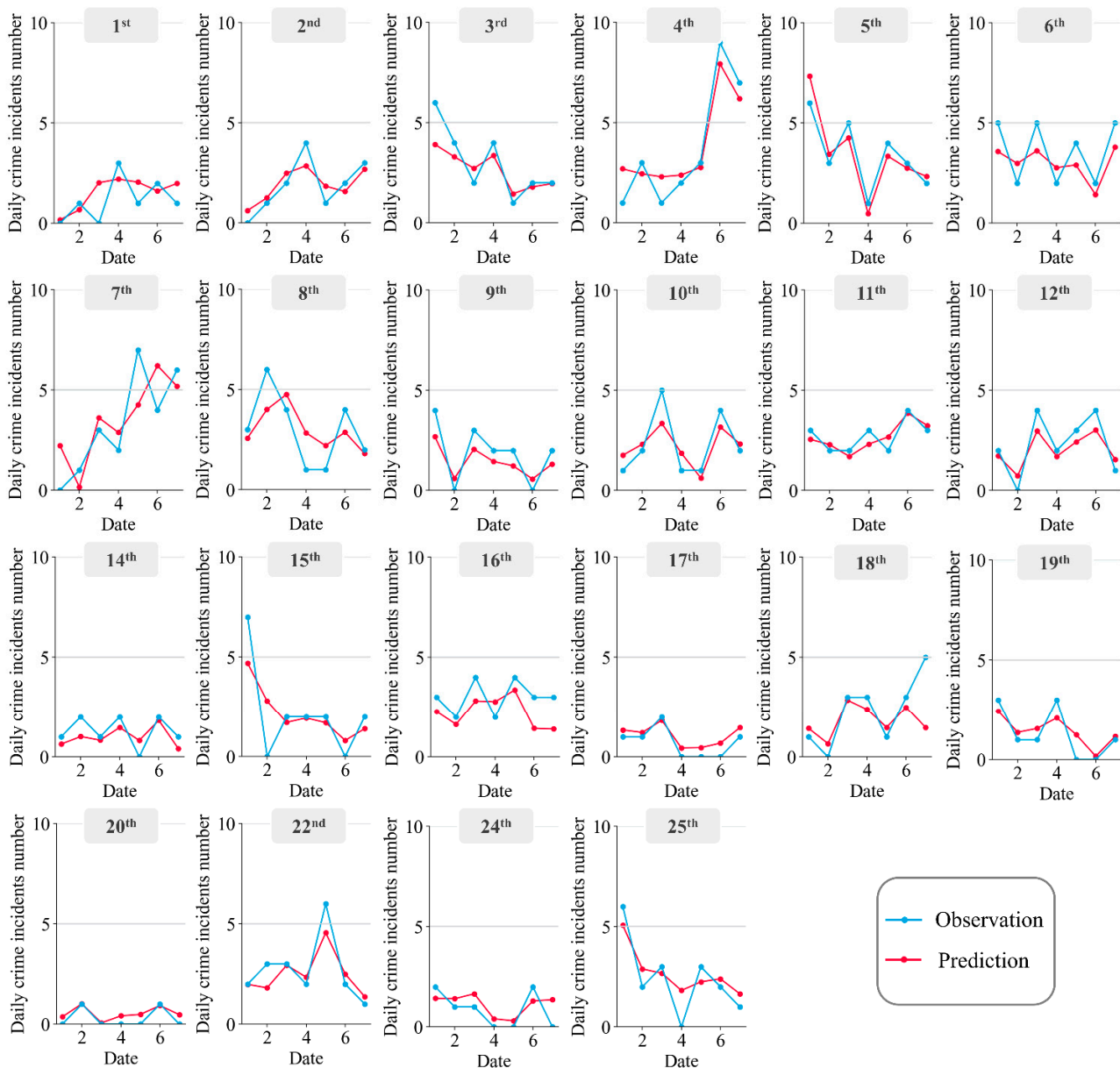
Table 3. Performances in spatial–temporal prediction of theft and assault in terms of MAE and RMSE.

Model	Theft		Assault	
	MAE	**RMSE**	**MAE**	**RMSE**
ARIMA	3.3145	3.8003	1.7594	2.1025
Ridge Regression	2.9168	3.4998	1.4917	1.9215
SVR	2.8030	3.3711	1.4609	1.8598
Random Forest	2.5064	2.9996	1.3191	1.5958
XGBoost	2.5114	3.0037	1.3218	1.5818
LSTM	1.8379	2.2395	1.1462	1.3778
CNN	1.4972	1.8079	0.8981	1.0819
Conv-LSTM	1.5574	1.8736	0.8269	0.9863
Our Model	1.3778	1.6318	0.7457	0.8851

4. Conclusions

In this study, an "online" integrated graph model based on an attention mechanism is proposed to predict the urban spatial–temporal crime patterns at a daily scale. This integrated graph model combines ResNet, GCN, and LSTM with the attention mechanism to extract and fuse the spatial–temporal features, topological graphs, and external features from the data of crime incidents. To validate this model, the daily theft and assault data at a police-district-scale of the city of Chicago in the US from 1 January 2015 to 7 January 2020 were employed. Indices MAE and RMSE were conducted to quantitatively evaluate this model, as well as eight other models (which were used as the comparisons) proposed in previous studies. The main findings of this study are summarized as follows:

(1) The predicted spatial–temporal patterns by using the proposed model have close agreements with the observations. Meanwhile, the average values of the two metrics (MAE and RMSE) are lower than two incidents for theft and one incident for assault, which indicates that this model achieves an extraordinarily high accuracy for predicting the spatial–temporal crime patterns. Moreover, the proposed model has a much better performance than all of the other eight models in terms of both MAE and RMSE. These results indicate that the integrated graph model can effectively predict the urban spatial–temporal distribution of crimes at a daily scale.

(2) This prediction method adopts irregular divisions on urban districts so that crime incident numbers in different police districts can be predicted. In addition, feature fusion is achieved by weighting the features through the attention mechanism, which provides an "online" framework to support the decision making for police, including patrolling and investigation, as well as other police resource allocation works.

The performance of this integrated graph model still has room for improvement. Our experimental results indicate that the prediction accuracy of this model would be lower in the districts where complicated social factors are aggregated. If more real-time and police-district-scale data, such as the house price, population density, traffic condition, and unemployment rate, are included in the future research, the performance of this model is expected to be further improved.

Author Contributions: Conceptualization, Xiaofeng Hu; methodology, Xiaofeng Hu, Miaomiao Hou, Xinge Han, Jitao Cai and Shuaiqi Yuan; validation, Xiaofeng Hu and Miaomiao Hou; formal analysis, Xiaofeng Hu and Miaomiao Hou; investigation, Xiaofeng Hu and Miaomiao Hou; resources, Xiaofeng Hu and Miaomiao Hou; data curation, Miaomiao Hou; writing—original draft preparation, Miaomiao Hou; writing—review and editing, Xiaofeng Hu, Miaomiao Hou, and Jitao Cai; visualization, Miaomiao Hou; supervision, Xiaofeng Hu; project administration, Xiaofeng Hu; funding acquisition, Xiaofeng Hu. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The data that support the findings of this study are available from the Chicago open data portal (https://data.cityofchicago.org/, accessed on 1 October 2020) and rp5 website (http://rp5.ru/, accessed on 1 October 2020).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. The descriptive statistics of theft and assault incidents numbers in the 22 police districts of Chicago from 1 January 2015 to 7 January 2020.

Police District	Theft					Assault				
	Mean	Standard deviation	Min	Max	Sum	Mean	Standard deviation	Min	Max	Sum
1st District	18.754	7.397	3	108	34,376	1.939	1.472	0	9	3554
2nd District	7.195	3.045	0	19	13,188	2.483	1.681	0	10	4552
3rd District	5.350	2.472	0	15	9806	3.011	1.798	0	9	5519
4th District	6.318	2.870	0	27	11,580	3.727	1.986	0	11	6831
5th District	4.801	2.327	0	14	8801	2.883	1.825	0	12	5285
6th District	8.178	3.172	0	26	14,991	3.982	2.130	0	12	7299
7th District	5.160	2.497	0	15	9458	3.552	2.036	0	12	6510
8th District	9.074	3.461	0	24	16,632	3.360	2.005	0	11	6158
9th District	6.155	2.798	0	18	11,283	2.591	1.702	0	10	4750
10th District	4.785	2.433	0	15	8771	2.586	1.683	0	10	4741
11th District	5.447	2.514	0	15	9985	3.594	2.025	0	13	6588
12th District	11.475	4.377	1	33	21,034	2.337	1.667	0	10	4283
14th District	9.362	3.775	0	25	17,160	1.460	1.268	0	7	2676
15th District	3.662	2.102	0	14	6712	2.346	1.583	0	9	4300
16th District	5.761	2.814	0	20	10,559	1.649	1.320	0	8	3023
17th District	5.706	2.703	0	16	10,459	1.199	1.133	0	6	2197
18th District	18.043	6.323	3	42	33,072	1.518	1.320	0	9	2782
19th District	11.890	4.442	1	47	21,795	1.536	1.251	0	7	2815
20th District	3.399	2.021	0	12	6231	0.803	0.903	0	5	1472
22nd District	4.687	2.470	0	17	8592	1.783	1.349	0	7	3269
24th District	5.451	2.674	0	17	9991	1.362	1.215	0	7	2497
25th District	7.389	2.998	0	18	13,544	2.809	1.751	0	11	5148
Total	168.041	28.187	61	309	308,020	52.509	10.937	23	88	96,249

Appendix B

The discomfort index (DI) is in line with the description of the human body temperature perception, which can be calculated according to temperature, humidity, and wind speed as follows:

$$DI = 0.5T_w + 0.5AT \tag{A1}$$

$$AT = 1.07T + 0.2e - 0.65v - 2.7 \tag{A2}$$

$$e = (RH \div 100) \times 6.105 \times e^{\frac{17.2T}{237.7+T}} \tag{A3}$$

$$T_w = AT\arctan(0.152\sqrt{RH + 8.314}) + \arctan(AT + RH)$$
$$-\arctan(RH - 1.676) + 0.0039RH^{\frac{3}{2}}\arctan(0.023RH) - 4.6803 \tag{A4}$$

where *AT* is the apparent temperature (°C), T_W is the thermodynamic wet-bulb temperature (°C), *T* is the air temperature (°C), *e* is the water vapor pressure (hPa), *V* is the wind speed (m/s), and *RH* is relative humidity (%). Weather data used to calculate DI in Chicago are obtained from the rp5 website (http://rp5.ru/, accessed on 1 October 2020).

References

1. Rummens, A.; Hardyns, W.; Pauwels, L. The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Appl. Geogr.* **2017**, *86*, 255–261. [CrossRef]
2. Kwon, E.; Jung, S.; Lee, J. Artificial Neural Network Model Development to Predict Theft Types in Consideration of Environmental Factors. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 99. [CrossRef]
3. Sherman, L.W.; Gartin, P.R.; Buerger, M.E. Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology* **1989**, *27*, 27–56. [CrossRef]

4. He, Z.; Tao, L.; Xie, Z.; Xu, C. Discovering spatial interaction patterns of near repeat crime by spatial association rules mining. *Sci. Rep.* **2020**, *10*, 17262. [CrossRef]

5. Gu, H.; Chen, P.; Li, H. Review and prospect of the research on the methods of crime space-time prediction. *J. Earth Inf. Sci.* **2021**, *23*, 43–57.

6. Lamari, Y.; Freskura, B.; Abdessamad, A.; Eichberg, S.; de Bonviller, S. Predicting Spatial Crime Occurrences through an Efficient Ensemble-Learning Model. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 645. [CrossRef]

7. Marques, S.C.; Ferreira, F.A.; Meidutė-Kavaliauskienė, I.; Banaitis, A. Classifying urban residential areas based on their exposure to crime: A constructivist approach. *Sustain. Cities Soc.* **2018**, *39*, 418–429. [CrossRef]

8. Maiti, A.; Zhang, Q.; Sannigrahi, S.; Pramanik, S.; Chakraborti, S.; Cerda, A.; Pilla, F. Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the contiguous United States. *Sustain. Cities Soc.* **2021**, *68*, 102784. [CrossRef]

9. Ingilevich, V.; Ivanov, S. Crime rate prediction in the urban environment using social factors. *Procedia Comput. Sci.* **2018**, *136*, 472–478. [CrossRef]

10. Yu, H.; Liu, L.; Yang, B.; Lan, M. Crime Prediction with Historical Crime and Movement Data of Potential Offenders Using a Spatio-Temporal Cokriging Method. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 732. [CrossRef]

11. Dash, S.K.; Safro, I.; Srinivasamurthy, R.S. Spatio-temporal prediction of crimes using network analytic approach. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 1912–1917.

12. Chen, P.; Hu, X.; Chen, J. The application of fuzzy information granulation and support vector machine in crime forecasting. *Sci. Technol. Eng.* **2015**, *35*, 54–57.

13. Pillai, R. *Optimized Predictive Modelling to Unfold the Links of Crime with Education, Safety and Climate in Chicago*; National College of Ireland: Dublin, Ireland, 2019.

14. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995.

15. Cheng, T.; Wang, J. Application of a dynamic recurrent neural network in spatio-temporal forecasting. In *Information Fusion and Geographic Information Systems*; Springer: St. Petersburg, Russia, 2007; pp. 173–186.

16. Li, W.; Wen, L.; Chen, Y. Spatial—Temporal forecast research of property crime under the driven of urban traffic factors. *Multimed. Tools Appl.* **2016**, *75*, 17669–17687.

17. Zhang, H.; Zhang, J.; Wang, Z.; Yin, H. An Adaptive Spatial Resolution Method Based on the ST-ResNet Model for Hourly Property Crime Prediction. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 314. [CrossRef]

18. Qian, Y.; Pan, L.; Wu, P.; Xia, Z. GeST: A Grid Embedding based Spatio-Temporal Correlation Model for Crime Prediction. In Proceedings of the 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), Hong Kong, China, 27–30 July 2020; pp. 1–7.

19. Zhang, T.; Ran, Y.; Wei, D. Application of Grid Management in Spatio-temporal Prediction of Crime. In Proceedings of the 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 22–24 May 2021; pp. 2745–2749.

20. Han, X.; Hu, X.; Wu, H.; Shen, B.; Wu, J. Risk Prediction of Theft Crimes in Urban Communities: An Integrated Model of LSTM and ST-GCN. *IEEE Access* **2020**, *8*, 217222–217230. [CrossRef]

21. Hu, T.; Zhu, X.; Duan, L.; Guo, W. Urban crime prediction based on spatio-temporal Bayesian model. *PLoS ONE* **2018**, *13*, e0206215. [CrossRef] [PubMed]

22. Yi, F.; Yu, Z.; Zhuang, F.; Zhang, X.; Xiong, H. An integrated model for crime prediction using temporal and spatial factors. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 1386–1391.

23. Sun, J.; Yue, M.; Lin, Z.; Yang, X.; Nocera, L.; Kahn, G.; Shahabi, C. CrimeForecaster: Crime Prediction by Exploiting the Geographical Neighborhoods' Spatiotemporal Dependencies. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Ghent, Belgium, 14–18 September 2020; pp. 52–67.

24. Chen, X.; Cho, Y.; Jang, S.Y. Crime prediction using Twitter sentiment and weather. In Proceedings of the 2015 Systems and Information Engineering Design Symposium, Charlottesville, VA, USA, 24–24 April 2015; pp. 63–68.

25. Zhao, X.; Tang, J. Modeling temporal-spatial correlations for crime prediction. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 497–506.

26. Wang, B.; Yin, P.; Bertozzi, A.L.; Brantingham, P.J.; Osher, S.J.; Xin, J. Deep learning for real-time crime forecasting and its ternarization. *Chin. Ann. Math. Ser. B* **2019**, *40*, 949–966. [CrossRef]

27. Hu, X.; Chen, P.; Huang, H.; Sun, T.; Li, D. Contrasting impacts of heat stress on violent and nonviolent robbery in Beijing, China. *Nat. Hazards* **2017**, *87*, 961–972. [CrossRef]

28. Hu, X.; Wu, J.; Chen, P.; Sun, T.; Li, D. Impact of climate variability and change on crime rates in Tangshan, China. *Sci. Total Environ.* **2017**, *609*, 1041–1048. [CrossRef]

29. Xu, R.; Xiong, X.; Abramson, M.J.; Li, S.; Guo, Y. Association between ambient temperature and sex offense: A case-crossover study in seven large US cities, 2007–2017. *Sustain. Cities Soc.* **2021**, *69*, 102828. [CrossRef]

30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

31. Loey, M.; Manogaran, G.; Taha, M.H.N.; Khalifa, N.E.M. Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. *Sustain. Cities Soc.* **2021**, *65*, 102600. [CrossRef] [PubMed]

32. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645.

33. Wang, B.; Zhang, D.; Zhang, D.; Brantingham, P.J.; Bertozzi, A.L. Deep learning for real time crime forecasting. *arXiv* **2017**, arXiv:1707.03340.

34. Lu, L.; Lyu, B. Reducing energy consumption of Neural Architecture Search: An inference latency prediction framework. *Sustain. Cities Soc.* **2021**, *67*, 102747. [CrossRef]

35. Wang, Y.; Jing, C. Spatiotemporal Graph Convolutional Network for Multi-Scale Traffic Forecasting. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 102. [CrossRef]

36. Bai, J.; Zhu, J.; Song, Y.; Zhao, L.; Hou, Z.; Du, R.; Li, H. A3t-gcn: Attention temporal graph convolutional network for traffic forecasting. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 485. [CrossRef]

37. Zhang, J.; Chen, F.; Guo, Y.; Li, X. Multi-graph convolutional network for short-term passenger flow forecasting in urban rail transit. *IET Intell. Transp. Syst.* **2020**, *14*, 1210–1217. [CrossRef]

38. Ma, X.; Tao, Z.; Wang, Y.; Yu, H.; Wang, Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transp. Res. Part C Emerg. Technol.* **2015**, *54*, 187–197. [CrossRef]

39. Yan, J.; Hou, M. Predicting Time Series of Theft Crimes Based on LSTM Network. *Data Anal. Knowl. Discov.* **2020**, *4*, 84–91.

40. Mao, W.; Wang, W.; Jiao, L.; Zhao, S.; Liu, A. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustain. Cities Soc.* **2021**, *65*, 102567. [CrossRef]

41. Treisman, A.M.; Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **1980**, *12*, 97–136. [CrossRef]

42. Li, Y.; Tong, Z.; Tong, S.; Westerdahl, D. A data-driven interval forecasting model for building energy prediction using attention-based LSTM and fuzzy information granulation. *Sustain. Cities Soc.* **2022**, *76*, 103481. [CrossRef]

43. Li, Y.; Zhu, Z.; Kong, D.; Han, H.; Zhao, Y. EA-LSTM: Evolutionary attention-based LSTM for time series prediction. *Knowl. Based Syst.* **2019**, *181*, 104785. [CrossRef]

44. Kim, S.; Kang, M. Financial series prediction using Attention LSTM. *arXiv* **2019**, arXiv:1902.10877.

45. Box, G.E.; Pierce, D.A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Am. Stat. Assoc.* **1970**, *65*, 1509–1526. [CrossRef]

46. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]

47. Awad, M.; Khanna, R. Support vector regression. In *Efficient Learning Machines*; Springer: Berkeley, CA, USA, 2015; pp. 67–80.

48. Alves, L.G.; Ribeiro, H.V.; Rodrigues, F.A. Crime prediction through urban metrics and statistical learning. *Phys. A Stat. Mech. Its Appl.* **2018**, *505*, 435–443. [CrossRef]

49. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H. Xgboost: Extreme gradient boosting. *R Package Version 0.4-2* **2015**, *1*, 1–4.

50. Cortez, B.; Carrera, B.; Kim, Y.-J.; Jung, J.-Y. An architecture for emergency event prediction using LSTM recurrent neural networks. *Expert Syst. Appl.* **2018**, *97*, 315–324. [CrossRef]

51. Duan, L.; Hu, T.; Cheng, E.; Zhu, J.; Gao, C. Deep convolutional neural networks for spatiotemporal crime prediction. In Proceedings of the International Conference on Information and Knowledge Engineering (IKE), Monte Carlo Resort, Las Vegas, NV, USA, 17–20 July 2017; pp. 61–67.

52. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.

53. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]

54. Shen, B.; Hu, X.; Wu, H. Impacts of climate variations on crime rates in Beijing, China. *Sci. Total Environ.* **2020**, *725*, 138190. [CrossRef]