

Article

Population Distribution Forecasting Based on the Fusion of Spatiotemporal Basic and External Features: A Case Study of Lujiazui Financial District

Xianzhou Cheng ^{1,2,3}, Xiaoming Wang ^{1,2,*} and Renhe Jiang ³ 

¹ Shanghai Earthquake Agency, Shanghai 200062, China; mn1472008@163.com

² Shanghai Sheshan National Geophysical, Shanghai 200062, China

³ Center for Spatial Information Science, The University of Tokyo, Tokyo 1138654, Japan; jiangrh@csis.u-tokyo.ac.jp

* Correspondence: wangyoucao@163.com



Citation: Cheng, X.; Wang, X.; Jiang, R. Population Distribution Forecasting Based on the Fusion of Spatiotemporal Basic and External Features: A Case Study of Lujiazui Financial District. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 395. <https://doi.org/10.3390/ijgi13110395>

Academic Editors: Wolfgang Kainz, Christos Chalkias, Marinos Kavouras, Margarita Kokla and Mara Nikolaidou

Received: 2 September 2024

Revised: 30 October 2024

Accepted: 3 November 2024

Published: 6 November 2024

Corrected: 27 January 2026

Abstract: Predicting the distribution of people in the time window approaching a disaster is crucial for post-disaster assistance activities and can be useful for evacuation route selection and shelter planning. However, two major limitations have not yet been addressed: (1) Most spatiotemporal prediction models incorporate spatiotemporal features either directly or indirectly, which results in high information redundancy in the parameters of the prediction model and low computational efficiency. (2) These models usually incorporate certain basic and external features, and they can neither change spatiotemporal addressed features according to spatiotemporal features nor change them in real-time according to spatiotemporal features. The spatiotemporal feature embedding methods for these models are inflexible and difficult to interpret. To overcome these problems, a lightweight population density distribution prediction framework that considers both basic and external spatiotemporal features is proposed. In the study, an autoencoder is used to extract spatiotemporal coded information to form a spatiotemporal attention mechanism, and basic and external spatiotemporal feature attention is fused by a fusion framework with learnable weights. The fused spatiotemporal attention is fused with Resnet as the prediction backbone network to predict the people distribution. Comparison and ablation experimental results show that the computational efficiency and interpretability of the prediction framework are improved by maximizing the scalability of the spatiotemporal features of the model by unleashing the scalability of the spatiotemporal features of the model while enhancing the interpretability of the spatiotemporal information as compared to the classical and popular spatiotemporal prediction frameworks. This study has a multiplier effect and provides a reference solution for predicting population distributions in similar regions around the globe.

Keywords: spatiotemporal fusion attention mechanism; lightweight framework; spatiotemporal feature interpretability



Copyright: © 2024 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Predictive models based on spatial-temporal sequences are widely employed in numerous domains, including healthcare, meteorology, transportation, and environmental protection [1–3]. Population distribution forecasting is the act of anticipating population density distribution data for future periods by creating a fitting model with historical data [4]. Real-time population density distribution is vital for the evacuation of people in the affected area. If you can precisely acquire the real-time flow of people, you may employ traffic control and evacuation plans to reduce casualties [5]. However, due to variables such as communication outages caused by disasters, collecting real-time people distribution after a disaster is challenging; therefore, population density distribution prediction modeling

becomes an alternative to real-time population density distribution [6]. The analysis of spatiotemporal characteristics is precious for the study of population forecasting and changes in the characteristics of population migration during sudden disasters such as earthquakes. By inferring the characteristics of population density distribution before and after sudden disaster events, we provide a better basis for understanding and analyzing the impact of sudden disaster events on population distribution [7]. The uncertainty and spatiotemporal connection in the distribution of geographic data components pose a significant difficulty in modeling the correct prediction of urban population density distribution using spatiotemporal data [8]. On the one hand, population dispersion is unevenly distributed over time; for example, the density of people in the early hours of the morning is substantially lower than throughout the day, and traffic conditions during the morning and evening rush hours will be similar during the week. However, the mechanisms of population dispersal vary geographically. For example, people are more likely to be concentrated in structurally dense areas than in vast natural spaces [9]. As a result, combining external and spatiotemporal variables to forecast population density distributions will make it easier to increase prediction accuracy [10].

Classical time series regression algorithms are frequently technically based and dominated by linear and nonlinear regression and average sliding kernels, which place major demands on the smoothness of linear regression models [11]. While deep-learning-type models with strong spatiotemporal data correlation extraction capabilities can typically significantly improve prediction performance, traditional methods that rely on descriptive methods using spatiotemporal dependencies are often ineffective in obtaining spatiotemporal feature correlations and thus providing high-quality predictions of spatial features [12]. Sophisticated methods for solving spatial–temporal geographic information models, including CA-LSTM [13] and ConvLSTM [14], have been made possible by advances in machine learning technology. As algorithms based on graph convolutional neural networks and graph generative models have obtained good results in the field of image and video prediction, efficient deep learning methods are beginning to be applied in the field of population density prediction [15]. While deep learning algorithms are capable of capturing the relationships between spatial–temporal sequential models, longer periods lead to worse predictive accuracy [16]. In contrast to image and language prediction, population density data are more likely to be significantly affected by external activities, and it is difficult for previous models to incorporate features of the external environment into the model features. However, spatiotemporal attention methods can incorporate external features into the model prediction through attentional fusion to further improve prediction performance.

As a lightweight spatiotemporal population density distribution prediction framework based on spatiotemporal feature attention, the framework utilizes spatiotemporal basic and external feature fusion to improve the prediction performance of pre-disaster population distribution. The spatiotemporal basic feature attention is obtained from the spatiotemporal feature dataset itself after compression and other processes, while the spatiotemporal external feature attention is obtained by having feature data that are closely related to the spatiotemporal feature data for embedding. The spatiotemporal basic and external feature attention mechanism is a deep learning model that combines spatiotemporal basic and external features. First, it abandons the mechanism of fusing multiple spatiotemporal features as inputs and instead weights the fusion through the attention mechanism to influence the backbone model's judgment of the importance of spatiotemporal features, thus simplifying the model while allowing for the incorporation of multiple external features to capture a broader range of spatiotemporal characteristics of the population distribution. Second, it uses a spatiotemporal autoencoder to nonlinearly activate spatiotemporal basic and external feature attention, allowing the model to flexibly learn the relevant weights within each spatiotemporal feature. Finally, a fusion framework with learnable parameters is used to weigh and fuse multiple spatiotemporal features, and the fused attention is incorporated into the backbone prediction network, which enables the backbone prediction network to enhance the spatiotemporal features related to the population distribution

characteristics and suppress the irrelevant spatiotemporal features through the learnable weights, thus improving the prediction accuracy and interpretability of the population density distribution. Our main contributions are specified below:

1. A lightweight and effective prediction framework is proposed, which may use basic and external spatiotemporal features to adequately capture the spatiotemporal variability of population density distribution data.
2. Using spatiotemporal attention to fuse basic and external spatiotemporal features rather than adding actual spatiotemporal feature data for population distribution prediction reduces the dimensionality of the predicted features, the model's parameters, and the complexity.
3. Controlling spatiotemporal feature pairings based on their respective feature weights enhances the model's interpretability while boosting the predictability of population density distribution.
4. Consistent with other population density prediction algorithms, it outperforms the baseline method in terms of prediction accuracy.

2. Related Works

Attentional mechanisms are effective for temporal feature fusion. SENet suggested a successful method of channel attention, yet it solely evaluated single-channel coding, ignoring the importance of spatiotemporal location information [17]. The attention mechanism has since developed on this basis and iterated rapidly toward efficient spatiotemporal multi-feature fusion [18].

2.1. Fusion of Attention Mechanisms

CBAM is an example of a hybrid attention mechanism that combines channel and spatial attention [19]. Efficient Channel Attention (ECA) simplifies the model by enabling cross-channel attentional interactions and has shown good results [20]. In contrast, Dual Attention Networks (DANs) [21] and Cross-Cutting Networks (CCNet) [22] use both non-local channel and non-local spatial attention to segment semantically. Hierarchical Attention Networks (HANs) [23] produce semantic similarity common attention graphs to align information about various modalities utilizing structural similarity. UFO-ViT eliminates the computational complexity of self-attention by modifying a few rows of self-attention to eliminate some nonlinearities. However, the approach is only applicable to dense data prediction [24]. Coordinate Attention [25] improves feature representation via perpendicular and parallel pooling to strengthen the connection between positional relations and channels. However, it is particularly suited for semantic segmentation obstacles.

2.2. Fusion of Temporal and Spatial Attention Mechanisms

Spatiotemporal feature fusion is the process of combining features from several spatiotemporal features, typically employing multiplication or addition for feature splicing [26]. The alignment process of spatiotemporal feature input data could constitute a bottleneck in the computational efficiency of spatiotemporal prediction models [27]. StNet [28] proposes a framework for integrating local and global spatiotemporal data via overlaid channels, but the model only supports shallow spatiotemporal fusion. The TCN-Transformer [29] parallel prediction model utilizes the Transformer to extract temporal features in sequences with long-term dependencies and utilizes a parallel structure to speed up model training and inference, but excessive attention to spatiotemporal early features triggers an imbalance in the attention mechanism. FusionFormer [30] implements a variable attention mechanism into the fusion coding module, which improves the flexibility of spatiotemporal fusion; nevertheless, the pre-processing of spatiotemporal comments in the early stages is overly sophisticated, increasing the model's complexity. AMGC [31] develops an innovative multi-graph attention mechanism that uses the inverse attention mechanism to fuse temporal dependencies from both global and local perspectives, but the model's over-reliance on graph attention in the spatial aspect increases the model parameters and reduces its

efficiency. ST-ResNet [32] fosters a residual convolution branch to model crowd traffic attributes and extracts time by convolving three types of features. Adding more spatiotemporal features increases the convolutional layers of the model, and the model's prediction efficiency decreases as the convolution parameters increase. ST-SSL1 [33], on the other hand, uses self-supervised learning to improve spatiotemporal features; however, increasing the number of auxiliary tasks improves prediction accuracy while increasing the model's computational complexity. Although STIN [34] uses a spatial-temporal aware convolutional layer instead of traditional convolution to improve the efficiency of capturing multimodal spatial-temporal information features, it is not appropriate with temporally discrete data owing to the use of RNN as the codec framework. With an increasing amount of fused channel features, another trend is to reduce the model structure while maintaining accuracy. To accelerate the training process and investigate deeper relationships between nodes, G-Fusion [35] accomplishes aggregation by randomly initializing the projection matrix. Residual attention maximizes the spatial attention of each object class while achieving good accuracy. At the same time, the computing cost has not increased considerably. MLCA [36] adds a modest number of parameters but dramatically improves accuracy. LSGA [37] extracts global deep semantic features using a hybrid spatial-spectral tagger rather than patch embedding and decreases model parameters by simplifying the embedding approach, resulting in good results.

Although multimodal spatial-temporal feature data can significantly improve the prediction ability of population distribution models, the direct embedding of multimodal data into the model results in a proliferation of model parameters, leading to redundancy and a decrease in computational efficiency. To address the dilemma of computational efficiency and model complexity, this research provides a lightweight and effective plug-and-play architecture. It enhances performance for population density distribution prediction and the interpretability of spatial-temporal aspects by drastically lowering model complexity.

3. Data

3.1. Study Area

As a research sample, we selected the population heat map data of Shanghai's Lujiazui Financial District. Located in the heart of Shanghai City, Lujiazui Financial District acts as the primary functional region of the Shanghai International Financial Center, host to the headquarters of numerous international financial organizations, and has one of Shanghai's densest populations. Lujiazui Financial District is a typical area of population distribution in Shanghai because of its high density of migrant workers, enormous inflow and outflow of population, and demographic characteristics that are more prone to changes in time and space (Figure 1).

3.2. Data Source

The heat map shows the distribution of the population, expressed as the density of the population per unit area. Heat map data are often used to study the dynamic distribution of people's characteristics in various types of cities and to analyze disaster avoidance strategies [38]. Heat maps are a useful tool for representing spatial-temporal series distributions, transforming huge volumes of data into informative visual summaries that do not require assumptions [39]. The source of our heat map data is Getui "<https://www.getui.com/>" (accessed on 27 June 2022), which is a major push technology service provider for Android and iOS apps. The heat map data we receive have been de-personalized and do not involve public privacy issues. To ensure the safety of private data, we use processed population heat map data, which lacks personal information.

We used demographic thermal data from 11 October 2021 to 26 June 2022, performed at three-hour intervals over a time spectrum, yielding a total of 2064 spatiotemporal data entries. Each spatiotemporal item of data contained latitude, longitude, and population density values for 570 sampling points (30×19), ranging from longitudes 121.48–121.53

to latitudes 31.22–31.25. The latitude and longitude data were recorded using the World Geographic Coordinate System (WGS-84).

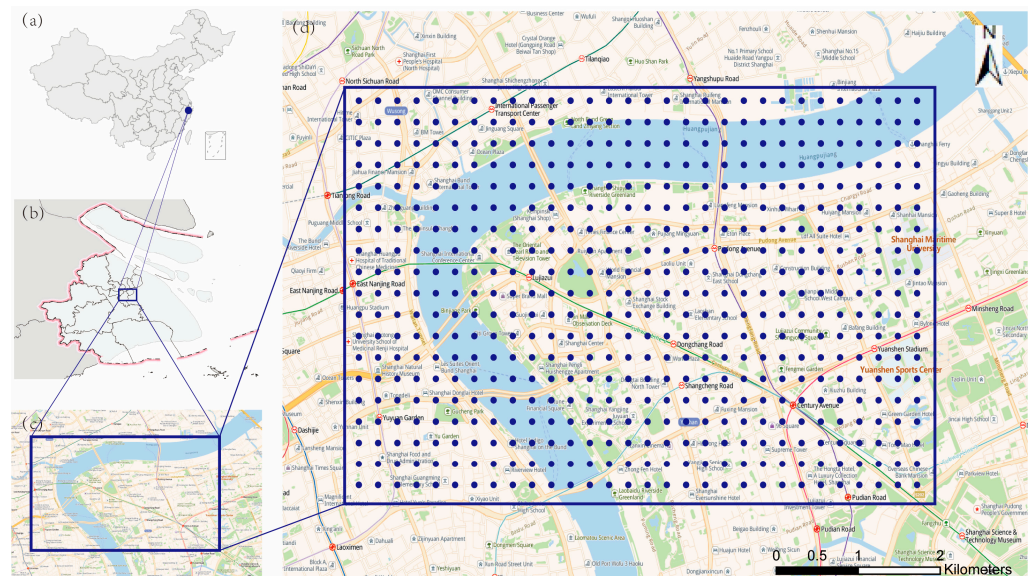


Figure 1. Distribution of the study area and sampling points: Lujiazui Financial District, Shanghai, with blue points indicating the sampled population density sampling area. (a) Location of the research area in China; (b) Location of the research area in China; (c) Scope of the study area; (d) Sampling points for population distribution.

The spatial–temporal external data were acquired from public information provided by the China Meteorological Administration (CMA) and the National Bureau of Surveying, Mapping, and Geographic Information (NBMSGI) and included data such as holidays, weather, wind, temperature, POI [40], and so on (Table 1).

Table 1. External temporal and spatial dataset.

Dataset	Type	Size
Holidays	1-D temporal data	258
Temperature (average)	1-D temporal data	258
Wind scale (12 levels)	1-D temporal data	258
Weather conditions (16 types)	1-D temporal data	258
Residential area (POI)	2-D spatial data	223
Shopping mall (POI)	2-D spatial data	1487
Restaurant (POI)	2-D spatial data	1249
Entertainment facilities (POI)	2-D spatial data	789

3.3. Data Framework

For the alignment and fusion of temporal and spatial data, in addition to guaranteeing the consistency of spatial–temporal feature attention, we need to develop a spatial–temporal data framework that matches basic and external spatial–temporal attention data. We construct the spatial–temporal attention frame components separately and configure the period window to θ time lengths of p cycles. The temporal features are then stacked and spliced along the time axis to obtain the frame sequence as: $F_t = [x_{t-l \times p}, x_{t-(l-1) \times p}, \dots, x_{t-p}]$, $x \in \mathbb{R}^{1 \times \theta}$, and the spatial features are then stacked and spliced along the time axis to produce the frame sequence as: $F_s = [x_{s-l \times p}, x_{s-(l-1) \times p}, \dots, x_{s-p}]$, $x \in \mathbb{R}^{H \times W \times \theta}$ (Figure 2).

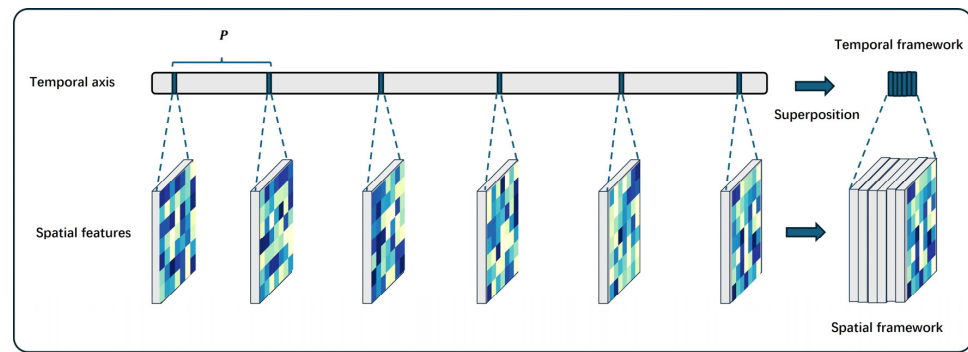


Figure 2. Schematic diagram of the spatial–temporal data framework.

3.4. Squeeze of SpatialTemporal Basic Data

To effectively compute temporal attention, we use the mean and maximum as base statistical compression functions to generate temporal base attention, which we then combine utilizing a variety of statistical compression functions, such as median and variance. After compressing the time domain features using time feature weights, learn the correlation of the time domain to time attention and extend it to the spatial domain to realize the time of the significant characteristics of the factor on the spatial domain features. Establishing a semantic response between the time domain and the spatial domain enhances the time domain’s semantic interpretation abilities about the global model. The time-domain basic x_{Basic}^T and additional squeezed features x_{Extra}^T are denoted as:

$$x_{Basic}^T = \sigma\{Avg[x(i, j)] + Max[x(i, j)]\}, i \in H, j \in W \quad (1)$$

$$x_{Extra}^T = \sigma\{\varphi_t[x(i, j)]\}, i \in H, j \in W, t \in M \quad (2)$$

where φ_t denotes the other statistical compression function of spatial–temporal features along the time direction, Avg and Max are the average and maximum pooling functions, the width of the spatiotemporal feature data is W, the length is H, and M is the number of temporal additional features. σ is the sigmoid function of activation.

Compared to the purpose of temporal feature acquisition, spatial feature attention is more inclined to emphasize the significance of “where”, or the significance of gaining spatial–temporal features in different locations [41]. The spatial domain base and additional feature compression are outlined as follows:

$$x_{Basic}^S = \sigma\{Avg[x(k)] + Max[x(k)]\}, k \in \theta \quad (3)$$

$$x_{Extra}^S = \sigma\{\varphi_s[x(k)]\}, k \in \theta, s \in N \quad (4)$$

where φ_s is denoted as the other statistical compression function of spatial–temporal features along the spatial direction, θ is the length of the time frame of spatial–temporal features, and N is the number of time-attached features. σ is the sigmoid function of activation.

3.5. Spatial–Temporal External Data Embedding

Many external temporal aspects can influence population distribution, and external conditions such as seasonal weather can also affect where people are distributed at different times and in various regions of the city [42]. We incorporate the effects of external features spatially and temporally using temporal and spatial frameworks. Let it be the feature vector for the external factors at prediction time t. To represent the effect of external features on prediction in our framework, we use the embedding method to convert external features into external feature attention, which complements the population density distribution task and improves the ability to characterize various spatiotemporal features.

The association of dynamics in temporal data is efficiently captured by mapping temporal features into feature vectors and storing them [43]. Temporal feature embedding

based on one-hot coding is independent of temporal features, whereas hash coding is a centralized method that maps the original features to a fixed-length vector space using a hash function [44]. The advantage of feature hashing is that it eliminates the need to record the mapping relationship between feature values and indexes, which can significantly reduce memory footprint. The feature hashing process can be stated in the following equation:

$$x_{Emb_t}^U = [W_{Emb_t}^1 x_t^1, \dots, W_{Emb_t}^T x_t^T], W_{Emb_t}^t \in \mathbb{R}^{1 \times \theta} \quad (5)$$

where x_t^1 is the original temporal feature, $W_{Emb_t}^T$ is the time frame weight function, $x_{Emb_t}^U$ is the external temporal feature after feature hash embedding, and U is the number of external temporal features.

The spatial external feature embedding is embedded into the spatial grid by the spatial features after the position transformation [45], columns such as POI point data $P(i, j)$ on the map (Figure 3), the mapping relationship of the spatial features in the grid position number N is as follows:

$$\Delta_i = |i_{max} - i_{min}| \quad (6)$$

$$\Delta_j = |j_{max} - j_{min}| \quad (7)$$

$$N_i = INT\left(\frac{i - i_{min}}{\Delta_i}\right) + 1 \quad (8)$$

$$N_j = INT\left(\frac{j - j_{min}}{\Delta_j}\right) + 1 \quad (9)$$

$$N_{ij} = INT\left(\frac{\Delta_i}{G_i}\right) \times N_j + N_i \quad (10)$$

$$x_{Emb_s}^V \leftarrow N_{ij}^V \quad (11)$$

where Δ_i, Δ_j denote the length and width of a single mapping grid, N_i, N_j denote the grid row and column numbers, N_{ij} denotes the absolute positional encoding of the data points in the grid, $x_{Emb_s}^V$ is the external temporal feature after grid embedding, V is the number of external spatial features, and G_i is the absolute position of the grid.

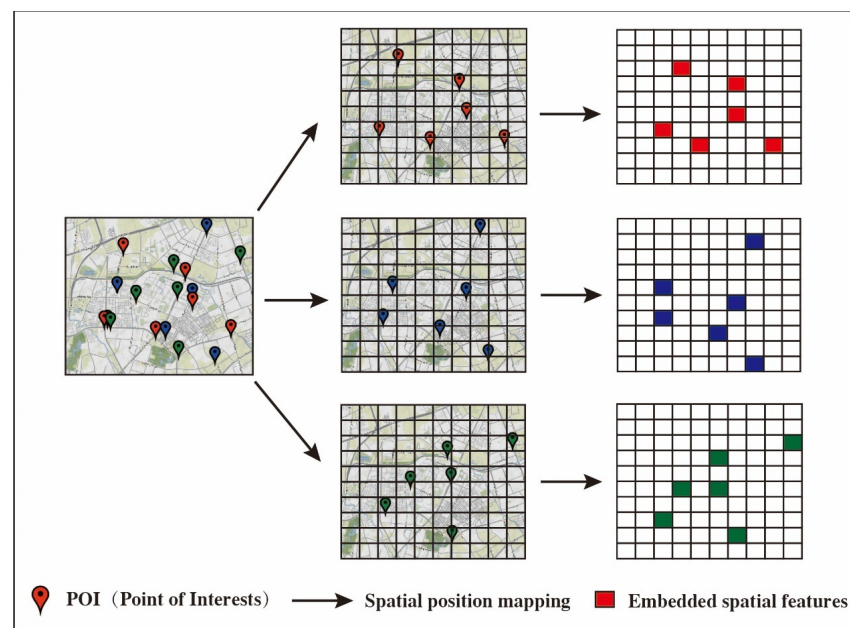


Figure 3. Spatial external feature embedding (POI point mapping).

4. Forecasting Framework

Spatial-temporal sequence modeling frequently necessitates a combination of temporal and spatial a priori knowledge and the adaptability of efficient and brief procedures and methods for generating accurate and dependable models and predictions [46].

In this section, we will examine how to integrate the spatial-temporal basic and external attention gained from the previous section's data architecture, as well as how to increase the interpretability of the basic and external spatial-temporal feature attention using learnable parameters. The prediction of population spatial-temporal distribution is a classic spatial-temporal serial regression problem. The population density tensor matrix for a specific time frame t is used as input, and the population density tensor matrix for the output moment $t + 1$ is predicted.

$$\hat{x}_{t+1} = \underset{\hat{x}_{t+1}}{\operatorname{argmax}} \rho(\hat{x}_{t+1} | x_{t-i+1}, \dots, x_t) \quad (12)$$

4.1. Temporal Feature Autoencoder

We construct a temporal autoencoder to encode temporal attention, adapt it to nonlinear correlations between distinct temporal features, and transmit temporal feature weights to temporal prediction mappings via attention [47]. We utilize a three-layer linear autoencoder as the temporal feature autoencoder. In this context, it is important to point out that the autoencoder's job is not to execute regression operations on temporal data but rather to autonomously learn the interrelationships between surrounding temporal characteristics, which does not necessitate the supervision of a decision model.

$$h_t = \sigma(x_t * W_{enc} + b_{enc}) \quad (13)$$

$$\tilde{x}_t = \sigma(h_t * W_{dec} + b_{dec}) \quad (14)$$

In the above equation, h is the encoding process of the temporal autoencoder, $\tilde{x}_t$ is the decoding process of the autoencoder, and W_{enc} and W_{dec} are the weights of the m-encoder and decoder, respectively. σ serves as an activation function that ensures that the temporal attention stays between 0 and 1.

4.2. Spatial Feature Autoencoder

In opposition to the linear autoencoder utilized for temporal feature attention, we embrace a convolutional autoencoder with convolution-transpose convolution as the foundation module to adequately capture spatial positional correlation weights [48]. We rebuild spatial features using a three-layer convolutional structure that replaces the autoencoder's fully connected layers with convolutional and transpose convolutional layers. This structure maximizes the efficiency of spatial feature weight learning while minimizing overlap. We input spatial features as x_s , the weight parameters of the convolutional encoder and the anti-convolutional decoder consist of W_{enc} and W_{dec} , and h_s is used to denote the context intermediate layer can be expressed as:

$$h_s = \sigma(x_s * W_{enc} + b_{enc}) \quad (15)$$

The parameters of the resulting convolutional layer are obtained by deconvolutional reconstruction:

$$\tilde{x}_s = \sigma(h_s * W_{dec} + b_{dec}) \quad (16)$$

4.3. Attention Fusion and Interpretation Module for Spatiotemporal Features

We reestablish spatial features using a three-layer convolutional structure that substitutes the autoencoder's fully connected layers with convolutional and transposed convolutional layers. This structure maximizes the efficiency of spatial feature weight learning while minimizing overlap [49]. Traditional spatial-temporal fusion prediction models capture spatial-temporal correlations using spatial-temporal topologies (for example, graph

convolution models), which has the advantage of completely exploiting spatial–temporal information to represent temporal dynamics and spatial dependencies. However, the issue is that utilizing models such as graph convolution might result in a large expansion of the number of covariates, making the calculation impractical, and an approximation method must be utilized to address this difficulty by reducing model accuracy [50]. To obtain the attention-weighted channel feature map, we first initialize the attention weights and then multiply them with each channel of the original feature map using autonomous learning. This will highlight channels that are useful for the current spatial–temporal prediction challenge while suppressing irrelevant channels [51].

$$\vec{w} = [w_0, w_1, \dots, w_K] \quad (17)$$

The weight vector after SoftMax normalization is as follows:

$$\hat{w} = \text{Softmax}(\vec{w})_i = \frac{e^{w_i}}{\sum_{j=1}^K e^{w_j}} \quad (18)$$

Temporal feature attention fusion is expressed as:

$$\text{Attention}_{Fusion}^T = \hat{w}_{Basic}^T * (x_t) + \hat{w}_{External}^T * (x_t) \quad (19)$$

Spatial feature attention fusion is expressed as:

$$\text{Attention}_{Fusion}^S = \hat{w}_{Basic}^S * (x_t) + \hat{w}_{External}^S * (x_t) \quad (20)$$

Secondly, we train and validate the model, and SoftMax normalizes the weight parameters obtained in the validation set to obtain the importance factor of each feature to obtain the spatial–temporal feature attention fusion and interpretation module.

$$\text{Attention}_{Fusion}^{ST} = \hat{w}_{Fusion}^T * (\text{Attention}_{Fusion}^T) * \hat{w}_{Fusion}^S * (\text{Attention}_{Fusion}^S) \quad (21)$$

$\vec{w}$ denotes the weight vector of the spatial–temporal feature attention plugin, σ denotes the sigmoid function, and w_i is the learnable weight.

4.4. Framework for Forecasting Population Distribution

The input features are denoted as $x \in \mathbb{R}^{H \times W \times T_{in}}$, where the height and width of the population heat map are H and W populations, respectively. The distributional forecasting backbone network can be represented as follows:

$$\hat{x} = \text{Res}(x) * \text{Attention}_{Fusion}^{ST} \quad (22)$$

where x and $\hat{x}$ are the inputs and outputs of the residual unit, respectively, and $\text{Res}()$ is the residual function that can be iteratively accumulated. Figure 4 depicts the overall structure of the spatial–temporal prediction framework, which comprises three components: the temporal basic and external attention module, the spatial basic and external attention module, and the spatial–temporal prediction module. We pool the temporal and spatial bases of features with length, width, and height of $w \times H \times T$. We derive the base temporal and spatial attention by pooling the features' temporal and spatial bases. Weighted fusion involves multiplying self-learned weights by each spatial–temporal component, followed by sigmoid activation for temporal and spatial attention. We feed the residuals from the population density tensor matrix with a time window of j -t into the ResNet backbone network, multiply them with the temporal and spatial fusion attention for weighted fusion, and then fuse them with the inputs to obtain the predicted spatial–temporal features at moment t .

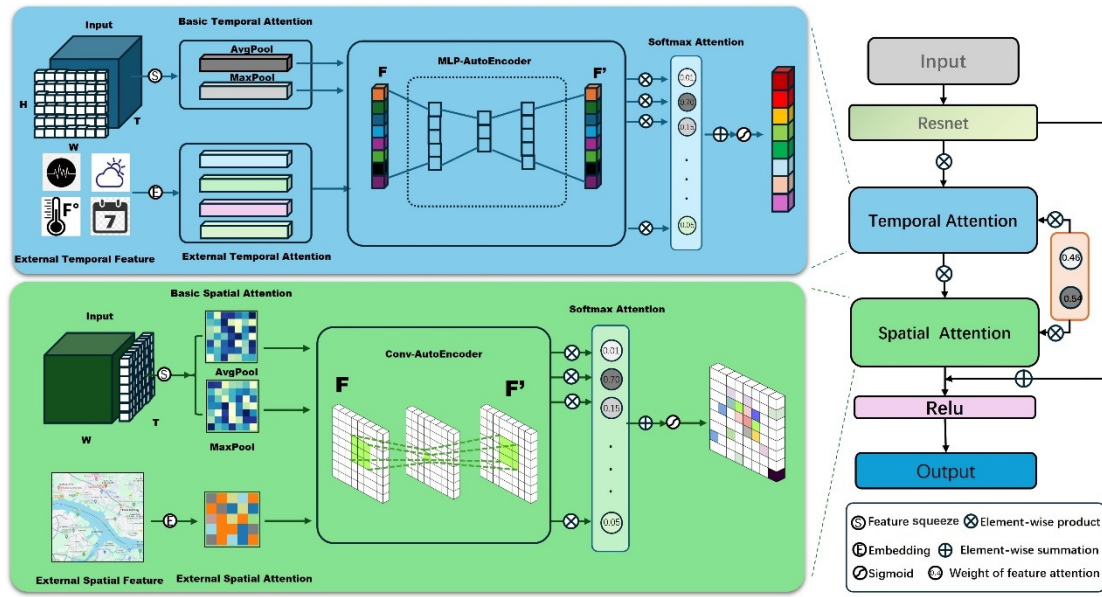


Figure 4. The framework for predicting the distribution of the attention population in and out of time and space.

The pseudocode of the prediction framework algorithm is released to visualize our model's basic structure and the algorithmic flow more intuitively. We added the spatiotemporal feature attribution method to the personnel distribution prediction model by means of an attention mechanism, which increases the transparency and interpretability of the model and avoids becoming a black box model (Algorithm 1).

Algorithm 1. Personnel density prediction algorithm

Input: Training dataset x_{train} , y_{train} ; External dataset x_{Extra}^T , x_{Extra}^S ; Predictive dataset $x_{predict}$, Minimum error ξ

Output: Result of training $\hat{x}_{train}$, Predicted results $\hat{x}_{predict}$

- 1 Initialize learnable basic spatial and temporal weights $\hat{w}_{Basic}^T$, $\hat{w}_{Basic}^S$
 - 2 Initialize learnable external spatial and temporal weights $\hat{w}_{External}^T$, $\hat{w}_{External}^S$
 - 3 Initialize learnable spatial–temporal fusion weights $\hat{w}_{Fusion}^T$, $\hat{w}_{Fusion}^S$
 - 4 **If** loss > ξ **Then:**
 - 5 **For** x_i, x_j, x_k in $x_{train}, x_{Extra}^T, x_{Extra}^S$:
 - 6 $x_{Basic}^T, x_{Basic}^S \leftarrow x_i$ by Equations (1) and (3)
 - 7 $x_{Emb_t}^U \leftarrow x_j$ by Equation (5)
 - 8 $x_{Emb_s}^U \leftarrow x_k$ by Equations (6)–(11)
 - 9 $\tilde{x}_t \leftarrow x_{Basic}^T, x_{Emb_t}^U$ by Equations (13) and (14)
 - 10 $\tilde{x}_s \leftarrow x_{Basic}^S, x_{Emb_s}^U$ by Equations (15) and (16)
 - 11 $Attention_{Fusion}^T \leftarrow \tilde{x}_t * \hat{w}_{Basic}^T + \tilde{x}_t * \hat{w}_{External}^T$
 - 12 $Attention_{Fusion}^S \leftarrow \tilde{x}_s * \hat{w}_{Basic}^S + \tilde{x}_s * \hat{w}_{External}^S$
 - 13 $\hat{x}_{train} \leftarrow Res(x) * \hat{w}_{Fusion}^T * Attention_{Fusion}^T * \hat{w}_{Fusion}^S * Attention_{Fusion}^S$
 - 14 Update $\hat{w}_{Basic}^T, \hat{w}_{Basic}^S, \hat{w}_{External}^T, \hat{w}_{External}^S, \hat{w}_{Fusion}^T, \hat{w}_{Fusion}^S$
 - 15 loss $\leftarrow$ MSELoss($\hat{x}_{train}, y_{train}$)
 - 16 **Else:**
 - 17 $\hat{x}_{predict} \leftarrow Res(x_{predict}) * \hat{w}_{Fusion}^T * Attention_{Fusion}^T * \hat{w}_{Fusion}^S * Attention_{Fusion}^S$
-

5. Experiments

The experimental software platform is RedHat Linux 6.7, the model training and testing environment is Python 3.6 and PyTorch 1.10, and the hardware system is CPU: 8 cores 4.0 GHz, memory 128 Gb, hard disk: 2 Tb, with 24 Gb video memory and a GPU computing card.

5.1. Experiment Parameters

We divided the population distribution data into a training set of 1840 data and a test set of 224 data. The basic feature window in both temporal and spatial domains is 56, the temporal external feature window is 8, the window sliding step is 1, the batch size is set to 8, the learning rate is set to 0.003, and the epoch is set to 100. The optimizer uses Adma. The framework training loss is Mse, and the framework validation parameters are parameters of the model computation speed and the accuracy of intersection and merger ratio. The temporal feature self-encoder is a three-layer linear layer, and the number of neurons in each layer is 8, 4, and 8, respectively. The spatial feature self-encoder consists of a convolution with a convolution kernel of 3 and a transposed convolution with the same convolution kernel of 3. The initial value of both temporal and spatial basic and external feature fusion weight training is 0.1667, and the total temporal and spatial feature fusion weight is 0.5.

5.2. Benchmark Models

We compare four typical spatiotemporal prediction models. ConvLSTM is a basic spatiotemporal long and short memory prediction model. PredRNN enhances spatiotemporal feature capture by stacking generation using ConvLSTM as a base module. STGCN describes the application of graphical convolution in spatiotemporal prediction. STTN predicts spatiotemporal features through encoder–decoder fusion and transformer fusion. Spatiotemporal features. The first two models reflect the classic framework of spatiotemporal feature prediction, while the latter two represent the popular framework of spatiotemporal feature prediction. We use classic frameworks as baselines to show the improvement of our prediction accuracy. We use currently popular frameworks as baselines to compare them, mainly to show the advantages of our model in terms of efficiency and interpretability.

- **ConvLSTM** We use a single-step prediction model of ConvLSTM with a 3×3 convolutional kernel, and the number of hidden state nodes in L is 64-32-16, respectively.
- **PreRNN** This model improves the ConvLSTM network by spatiotemporal memory flow (M). We use 4 ST-LSTMs with 128 hidden states per layer.
- **STGCN** The underlying architecture is a spatial convolution sandwiched between two temporal convolutions. We used two ST blocks with a kernel size of 3×3 .
- **STTN** is a transformer-based spatiotemporal prediction model with high prediction accuracy and consistency. The 2D convolutional kernel of the generator is 3×3 . The kernel for the discriminator 3D convolution is 3.

5.3. Experiments Result

When compared to other techniques, the spatiotemporal basic and external feature fusion framework is more effective at predicting population distribution. The significant reduction in model parameters and increase in prediction speed is primarily due to the frame-lightening work of the spatiotemporal basic and external feature fusion framework employing spatiotemporal attention, i.e., fusing the compressed attention of basic and external features rather than the actual features. The rise in prediction accuracy and intersection ratio, on the other hand, is due to the framework's increased spatiotemporal perception via multi-feature fusion (Table 2).

Figure 5 shows that the anticipated population distribution of ConvLSTM, PredRNN, STGCN, and STTN loses a lot of detail. Our system recognizes details more accurately than other models. On the other hand, our network's multi-temporal feature fusion function enhances its ability to capture details. Even when compared to STTN, which has strong

prediction performance, our model detects deeper and more precise features. Compared to STTN, our model predicts better at night, which highlights the advantage of attention to spatiotemporal basic and external features to maximize the capture of the spatiotemporal characteristics of the population distribution. Our model predicts far better than the other three models, both in the dark and in the daytime. This suggests that the spatiotemporal basic and external attention mechanism has an advantage over the four models mentioned above in the problem of people distribution prediction.

Table 2. Comparison with benchmark model. The up arrow indicates that the larger the parameters, the better the model performance, while the down arrow indicates that the smaller the parameters, the better the model performance.

Models	Parameters (M) ↓	Speed (s/epoch) ↓	Acc (%) ↑	Mse ↓	mIoU ↑
ConvLSTM	126.07	15.17	63.31	0.339	69.51
PredRNN	175.28	36.55	65.80	0.216	73.62
STGCN	240.78	18.58	80.17	0.077	81.38
STTN	302.85	51.70	82.48	0.051	84.91
Ours	43.52	7.31	97.07	0.042	86.95

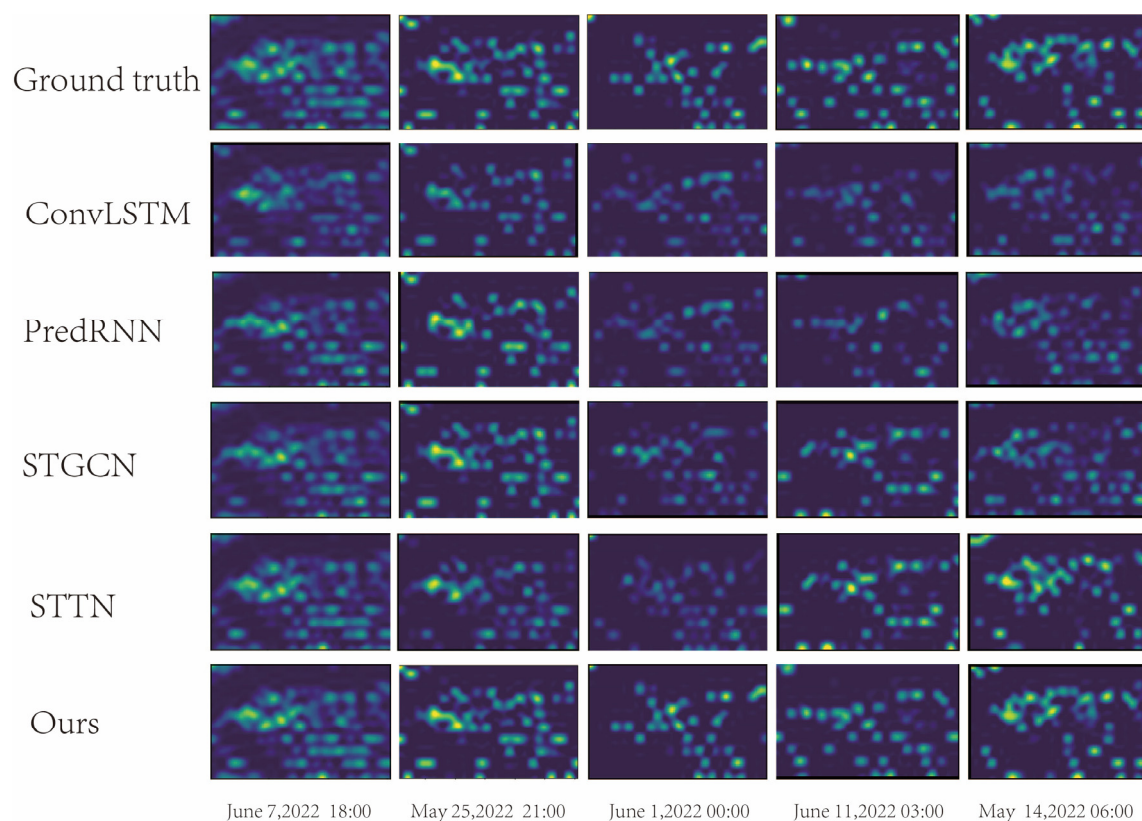


Figure 5. Comparison of selected 5-time window prediction results with other models.

Figures 6 and 7 show the weights of temporal basic and external feature attention, in addition to spatial basic and external feature attention. Based on the figure, we draw three conclusions:

1. The attentional weights of basic spatiotemporal features dominated the population distributional characteristics. The attentional weights of external spatiotemporal features were lower than the attentional weights of basic spatiotemporal features most of the time, and the distributions of the weights proved that the spatiotemporal features

themselves played a primary role in predicting the distribution of the population and the basic and external spatiotemporal features.

2. In the prediction process, the weights of temporal and spatial variables are comparable, implying that temporal and spatial features contributed equally to the population distribution prediction problem.
3. External spatiotemporal features have a low weight, yet they are critical in complementing the spatiotemporal basis features and boosting the prediction framework's interpretability.

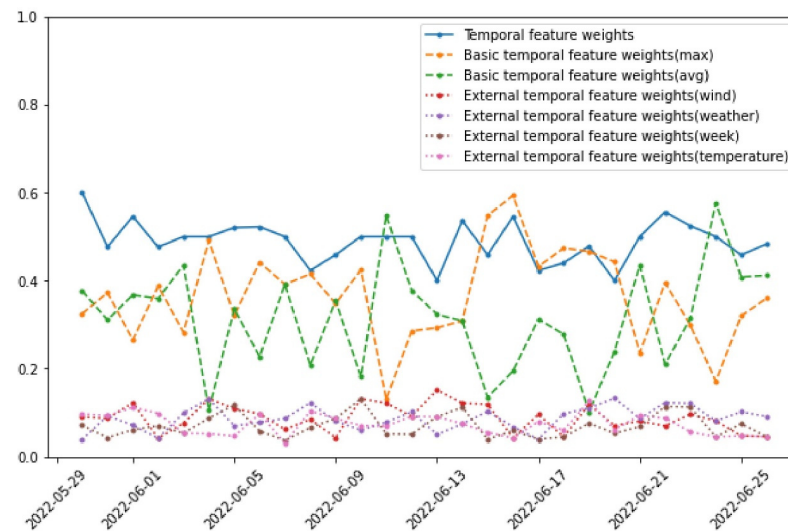


Figure 6. Attentional fusion weights of temporal basic and external features for test data.

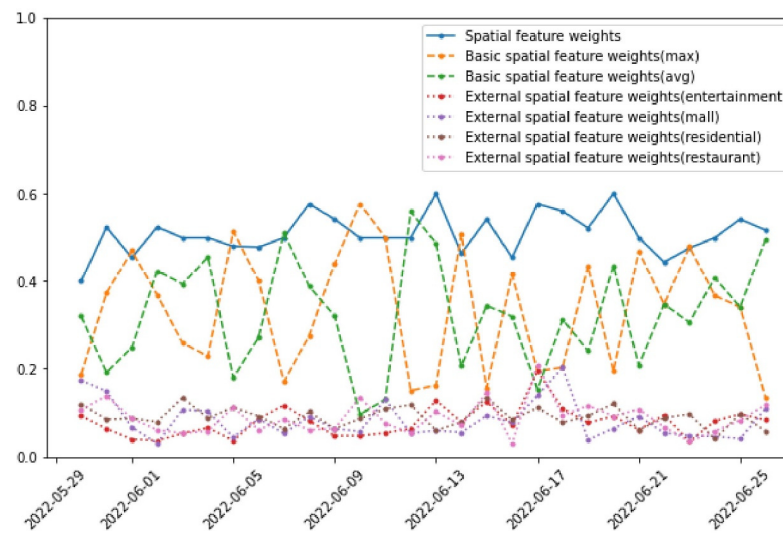


Figure 7. Attentional fusion weights of spatial basic and external features for test data.

5.4. Ablation Experiments

We employ ResNet-18, 34, and 50 as the foundation for ablation experiments on spatiotemporal feature attention models, respectively.

To validate the framework's effectiveness, we evaluate our model using a combination of the mean square error (MSE), mean average error (MAE), root mean square error (RMSE), the ratio of the regression sum of squares to the total sum of squares (R²), and peak signal-

to-noise ratio. The prediction improves when the MSE, RMSE, and MAE decrease and the R2 and PSNR increase. The performance measurement formulas are as follows:

$$MSE = \frac{1}{n} \sum_{t=1}^n (\hat{x}_t - x_t)^2 \quad (23)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (\hat{x}_t - x_t)^2} \quad (24)$$

$$MAE = \frac{1}{n} \sum_{t=1}^n |\hat{x}_t - x_t| \quad (25)$$

$$R^2 = 1 - \frac{\sum_{t=1}^n (\hat{x}_t - x_t)^2}{\sum_{t=1}^n (\hat{x}_t - \bar{x})^2} \quad (26)$$

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) = 20 \cdot \log_{10} \left(\frac{MAX_I}{\sqrt{MSE}} \right) \quad (27)$$

where $\hat{x}_t$ denotes the predicted value of the population distribution at moment t and x_t is the true value of the population distribution at moment t . n is the total number of prediction windows.

As illustrated in Table 3, T stands for temporal attention, S for spatial attention, ST for spatiotemporal attention, and STE for spatiotemporal basic and external attention. The loss of the model decreases insignificantly as the depth of the backbone network increases, whereas the loss decreases with the gradual increase in the spatiotemporal basic and external feature attention, both of which confirm that the spatiotemporal basic and external feature fusion module is critical to the prediction results.

Table 3. Results of an ablation study of spatiotemporal feature attention fusion frame. The up arrow indicates that the larger the parameters, the better the model performance, while the down arrow indicates that the smaller the parameters, the better the model performance.

Backbone	Attention	MSE ↓	MAE ↓	RMSE ↓	R ² ↑	PSNR ↑
ResNet-18		0.052	0.107	0.225	0.955	61.34
	+T	0.051	0.091	0.224	0.963	61.28
	+S	0.051	0.091	0.218	0.962	61.40
	+ST	0.046	0.090	0.216	0.967	61.39
	+STE	0.045	0.089	0.214	0.972	61.64
ResNet-34		0.049	0.103	0.220	0.950	61.04
	+T	0.049	0.093	0.220	0.961	61.35
	+S	0.053	0.095	0.229	0.960	61.00
	+ST	0.046	0.093	0.216	0.963	61.95
	+STE	0.043	0.086	0.206	0.973	62.05
ResNet-50		0.053	0.102	0.230	0.952	61.08
	+T	0.053	0.098	0.228	0.968	61.03
	+S	0.059	0.103	0.242	0.959	60.48
	+ST	0.048	0.097	0.217	0.971	61.55
	+STE	0.042	0.085	0.205	0.973	61.97

6. Discussion

Population distribution data are a typical spatiotemporal dynamic dataset characterized by real-time, spatiotemporal heterogeneity, and spatiotemporal dependence. Spatiotemporal feature fusion can take advantage of the characteristics of spatiotemporal

dynamic data to significantly improve the model's insight into spatiotemporal dynamic features, thus improving the performance of predictive models. Therefore, the spatiotemporal feature fusion mechanism has rapidly become a current academic hotspot and is widely used in spatiotemporal data prediction. We innovatively propose an efficient spatiotemporal feature fusion scheme that can significantly reduce the computational overhead of model training while improving the accuracy of population distribution prediction. Instead of the traditional direct fusion of spatiotemporal features, we use a weighted fusion of spatiotemporal attention mechanisms to predict population distribution, which improves computational speed and prediction accuracy compared with the traditional fusion model. By fusing spatiotemporal basic and external spatiotemporal features with real-time learnable weights, we can not only dynamically fit the population distribution law but also utilize the real-time change in weights to improve the dynamic interpretability of the prediction model.

Comparison experiments were conducted using four baseline spatiotemporal prediction frameworks, and thanks to the spatiotemporal feature attention fusion mechanism we employed, our model has the smallest number of parameters and the fastest computation speed. However, the computational accuracy parameter is the highest, and the final experiments also show that our results are closer to the true values. This indicates that the computational efficiency and accuracy of our model are the best. The demonstration of the dynamics of the real-time spatiotemporal feature fusion weight changes shows the excellent interpretability of our model. In addition, our framework can fuse many different forms of internal and external spatiotemporal feature data without changing the structure of the model framework, which has good scalability. To test whether the components of the framework are effective, we conducted ablation experiments on the framework's backbone, temporal attention, spatial attention, spatiotemporal attention, and inter-temporal attention. The results demonstrate that spatial interval attention is the most effective of all components, ensuring that spatial interval and spatial interval fusion of attention play a decisive role in the accuracy of population distribution prediction.

Firstly, we fuse basic and external spatiotemporal features in our framework, which fully exploits the complementary advantages between different spatiotemporal modal data and makes up for the deficiencies in the spatiotemporal features of a single prediction data to better reflect the spatiotemporal heterogeneity characteristics of population distribution. Secondly, to avoid the huge computational overhead brought by the direct fusion of spatiotemporal features, we utilize the lightweight spatiotemporal attention fusion mechanism, which can fuse multiple spatiotemporal features to capture spatiotemporal and spatial internal and external features as much as possible without increasing the computational burden of the backbone network, to better capture the spatiotemporal dependence of the population distribution and improve the prediction performance. Thirdly, the fusion of rich spatiotemporal features is the basis for improving the accuracy of population distribution prediction, and the learnable weights and spatiotemporal autoencoder are the keys to improving the accuracy of population distribution prediction. The variation in spatiotemporal feature fusion weights then demonstrates the dynamic interpretability of the model. Finally, compared with the basic population distribution prediction method, our model achieves higher prediction accuracy in three aspects: (1) richer spatiotemporal features provide data security; (2) nonlinear weights and spatiotemporal autoencoder provide flexibility to capture spatiotemporal features in real time for spatiotemporal prediction; and (3) efficiently spatiotemporal feature attentional fusion mechanisms can increase the number of strongly correlated spatiotemporal features and decrease the number of weakly correlated spatiotemporal features in real-time. Spatiotemporal features in the prediction model.

However, we have not taken into account the rapid short-term sudden-onset features in the spatiotemporal external characteristics, and thus the usefulness of our model in sudden-onset disasters is deficient. In particular, due to the large time window span of our population distribution data, it is not conducive to capturing population movement and implied population density during sudden-onset disasters. Thus, our model is not

highly applicable to the problem of population distribution prediction under fast-onset disasters, but our prediction model can be used with caution under disasters with long time spans. Our next step will be to study the importance of each weight for the global spatial and temporal aspects, adding the population distribution data of sudden-onset disaster events to further study the model and strive to accurately predict the distribution of the population within the time window of the disaster and exclude the interference and influence of other factors. The characteristics of the impact of each disaster on the distribution of population are analyzed. In addition, we will extend our framework and embed long- and short-term emergencies, such as earthquakes, fires, and epidemics, in subsequent work. Find data with denser time windows for forecasting and optimize our model to better capture population movement patterns during major events.

7. Conclusions

We investigate the weighting of spatiotemporal feature attention, the primary and secondary links between basic spatiotemporal features and external features, and the importance of internal and external spatiotemporal attention in predicting population distributions, and propose a novel framework for predicting population distributions in a lightweight manner. The framework improves the ability to capture spatiotemporal population distribution data by weighted fusion of basic spatiotemporal information and external spatiotemporal features. Our study area is the Lujiazui financial district, a typical population distribution area in Shanghai. The population distribution in this area has a high degree of spatiotemporal dependence, and the spatiotemporal features are numerous and complex. Experiments show that our prediction model can fully integrate various spatiotemporal features for fast and accurate prediction of population distribution under low computing power, which also provides targeted research and solutions for similar spatiotemporally sensitive population distribution cities around the world. In response to the problem of population data disruption caused by sudden disaster events, our model can compensate for the lack of pre-disaster data by reliably and efficiently predicting population distribution results that are similar to the pre-disaster data, which will effectively help rescuers plan evacuation routes and people in the disaster area to choose the nearest shelter for real-time risk avoidance. These data can be compared and analyzed with the post-disaster population distribution data to verify whether the disaster shelter meets the needs of the population in the area and plays a theoretical role in supporting the evaluation and re-planning of the post-disaster shelter. The spatiotemporal dynamic distribution characteristics of the population have important scientific value and practical significance for urban disaster emergency response and pre-disaster preparation. We utilize the spatiotemporal dependence of the population distribution of a typical city, weight and integrate all kinds of spatiotemporal basic and external characteristics, and use a small number of computational resources to improve the spatiotemporal prediction accuracy and efficiency of the population distribution characteristics, thus providing real-time population distribution data for the evacuation plan, escape routes, and shelter selection of the disaster preparedness and disaster response. This provides real-time population distribution data support for evacuation plans, escape routes, and shelter locations for disaster preparedness and response, enabling people to evacuate in an orderly manner during emergencies, minimizing casualties, and helping people to respond in time to reduce losses caused by disasters.

Author Contributions: Conceptualization, Xianzhou Cheng; methodology, Xianzhou Cheng; software, Xianzhou Cheng; validation, Xianzhou Cheng; formal analysis, Xianzhou Cheng; resources, Xiaoming Wang; data curation, Xianzhou Cheng; writing—original draft preparation, Xianzhou Cheng; writing—review and editing, Renhe Jiang; visualization, Xianzhou Cheng; supervision, Xiaoming Wang; project administration, Xiaoming Wang; funding acquisition, Xiaoming Wang. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The authors do not have permission to share data.

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

References

1. Du, Y.; Xu, Y.; Wang, X.; Liu, L.; Ma, P. EEG Temporal–Spatial Transformer for Person Identification. *Sci. Rep.* **2022**, *12*, 14378. [[CrossRef](#)] [[PubMed](#)]
2. Varshney, Y.; Kumar, V.; Dubey, D.K.; Sharma, S. Forecasting Precision: The Role of Graph Neural Networks and Dynamic GNNs in Weather Prediction. *J. Big Data Technol. Bus. Anal.* **2024**, *3*, 28–33.
3. Oliveira Santos, V.; Costa Rocha, P.A.; Scott, J.; Van Griensven Thé, J.; Gharabaghi, B. Spatiotemporal Air Pollution Forecasting in Houston-TX: A Case Study for Ozone Using Deep Graph Neural Networks. *Atmosphere* **2023**, *14*, 308. [[CrossRef](#)]
4. Song, X.; Zhang, Q.; Sekimoto, Y.; Shibasaki, R. Prediction of Human Emergency Behavior and Their Mobility Following Large-Scale Disaster. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; ACM: New York, NY, USA, 2014; pp. 5–14.
5. Wei, B.; Nie, G.; Su, G.; Sun, L.; Bai, X.; Qi, W. Risk Assessment of People Trapped in Earthquake Based on Km Grid: A Case Study of the 2014 Ludian Earthquake, China. *Geomat. Nat. Hazards Risk* **2017**, *8*, 1289–1305. [[CrossRef](#)]
6. Ara, S. Impact of Temporal Population Distribution on Earthquake Loss Estimation: A Case Study on Sylhet, Bangladesh. *Int. J. Disaster Risk Sci.* **2014**, *5*, 296–312. [[CrossRef](#)]
7. Wei, B.; Hu, B.; Qi, W. Fine-Scale Spatiotemporal Distribution Assessment of Indoor Population Based on Single Buildings: A Case in Dongcheng Subdistrict, Xichang, China. *Sustainability* **2023**, *15*, 7423. [[CrossRef](#)]
8. Jiang, R.; Cai, Z.; Wang, Z.; Yang, C.; Fan, Z.; Chen, Q.; Tsubouchi, K.; Song, X.; Shibasaki, R. DeepCrowd: A Deep Model for Large-Scale Citywide Crowd Density and Flow Prediction. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 276–290. [[CrossRef](#)]
9. Li, J.; Li, J.; Yuan, Y.; Li, G. Spatiotemporal Distribution Characteristics and Mechanism Analysis of Urban Population Density: A Case of Xi'an, Shaanxi, China. *Cities* **2019**, *86*, 62–70. [[CrossRef](#)]
10. Jiang, R.; Wang, Z.; Yong, J.; Jeph, P.; Chen, Q.; Kobayashi, Y.; Song, X.; Fukushima, S.; Suzumura, T. Spatio-Temporal Meta-Graph Learning for Traffic Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 8078–8086. [[CrossRef](#)]
11. Swain, S.; Nandi, S.; Patel, P. Development of an ARIMA Model for Monthly Rainfall Forecasting over Khordha District, Odisha, India. In *Recent Findings in Intelligent Computing Techniques*; Sa, P.K., Bakshi, S., Hatzilygeroudis, I.K., Sahoo, M.N., Eds.; Advances in Intelligent Systems and Computing; Springer Singapore: Singapore, 2018; Volume 708, pp. 325–331, ISBN 978-981-10-8635-9.
12. Wang, S.; Cao, J.; Yu, P.S. Deep Learning for Spatio-Temporal Data Mining: A Survey. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 3681–3700. [[CrossRef](#)]
13. Du, C.; Shu, P.; Li, Y. CA-LSTM: Search Task Identification with Context Attention Based LSTM. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; ACM: Ann Arbor, MI, USA, 2018; pp. 1101–1104.
14. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.; Woo, W. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015.
15. Wang, P.; Huang, X.; Mango, J.; Zhang, D.; Xu, D.; Li, X. A Hybrid Population Distribution Prediction Approach Integrating LSTM and CA Models with Micro-Spatiotemporal Granularity: A Case Study of Chongming District, Shanghai. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 544. [[CrossRef](#)]
16. Zhang, S.; Zhang, J.; Yang, L.; Yin, J.; Gao, Z. Spatiotemporal Attention Fusion Network for Short-Term Passenger Flow Prediction on New Year's Day Holiday in Urban Rail Transit System. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 59–77. [[CrossRef](#)]
17. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
18. Wang, Z.; Xia, T.; Jiang, R.; Liu, X.; Kim, K.-S.; Song, X.; Shibasaki, R. Forecasting Ambulance Demand with Profiled Human Mobility via Heterogeneous Multi-Graph Neural Networks. In Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 19–22 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1751–1762.
19. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19, ISBN 978-3-030-01233-5.
20. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 11531–11539.
21. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3141–3149.
22. Mutz, D.C. The Consequences of Cross-Cutting Networks for Political Participation. *Am. J. Polit. Sci.* **2002**, *46*, 838. [[CrossRef](#)]

23. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 22–27 April 2007; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 1480–1489.
24. Song, J. UFO-ViT: High Performance Linear Vision Transformer without Softmax. *arXiv* **2021**, arXiv:2109.14382.
25. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 13708–13717.
26. Ma, H.; Zhou, M.; Ouyang, X.; Yin, D.; Jiang, R.; Song, X. Forecasting Regional Multimodal Transportation Demand with Graph Neural Networks: An Open Dataset. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 3263–3268.
27. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional Feature Fusion. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3559–3568.
28. He, D.; Zhou, Z.; Gan, C.; Li, F.; Liu, X.; Li, Y.; Wang, L.; Wen, S. StNet: Local and Global Spatial-Temporal Modeling for Action Recognition. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 8401–8408. [[CrossRef](#)]
29. Chen, H.; Tian, A.; Zhang, Y.; Liu, Y. Early Time Series Classification Using TCN-Transformer. In Proceedings of the 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Dali, China, 12–14 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1079–1082.
30. Hu, C.; Zheng, H.; Li, K.; Xu, J.; Mao, W.; Luo, M.; Wang, L.; Chen, M.; Peng, Q.; Liu, K.; et al. FusionFormer: A Multi-Sensory Fusion in Bird’s-Eye-View and Temporal Consistent Transformer for 3D Object Detection. *arXiv* **2023**, arXiv:2309.05257.
31. Luo, Y.; Liu, Q.; Zhu, H.; Fan, H.; Song, T.; Yu, C.W.; Du, B. Multistep Flow Prediction on Car-Sharing Systems: A Multi-Graph Convolutional Neural Network with Attention Mechanism. *Int. J. Softw. Eng. Knowl. Eng.* **2019**, *29*, 1727–1740.
32. Zhang, J.; Zheng, Y.; Qi, D. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*. [[CrossRef](#)]
33. Ji, J.; Wang, J.; Huang, C.; Wu, J.; Xu, B.; Wu, Z.; Zhang, J.; Zheng, Y. Spatio-Temporal Self-Supervised Learning for Traffic Flow Prediction. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 4356–4364. [[CrossRef](#)]
34. Jin, Q.; Zhang, X.; Xiao, X.; Wang, Y.; Meng, G.; Xiang, S.; Pan, C. SpatioTemporal Inference Network for Precipitation Nowcasting With Multimodal Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 1299–1314. [[CrossRef](#)]
35. Liu, S.; Yao, S.; Li, J.; Liu, D.; Wang, T.; Shao, H.; Abdelzaher, T. GlobalFusion: A Global Attentional Deep Learning Framework for Multisensor Information Fusion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *4*, 1–27. [[CrossRef](#)]
36. Wan, D.; Lu, R.; Shen, S.; Xu, T.; Lang, X.; Ren, Z. Mixed Local Channel Attention for Object Detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106442. [[CrossRef](#)]
37. Ma, C.; Wan, M.; Wu, J.; Kong, X.; Shao, A.; Wang, F.; Chen, Q.; Gu, G. Light Self-Gaussian-Attention Vision Transformer for Hyperspectral Image Classification. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–12. [[CrossRef](#)]
38. Feng, D.; Tu, L.; Sun, Z. Research on Population Spatiotemporal Aggregation Characteristics of a Small City: A Case Study on Shehong County Based on Baidu Heat Maps. *Sustainability* **2019**, *11*, 6276. [[CrossRef](#)]
39. Storme, T.; Derudder, B.; Dörny, S. Introducing Cluster Heatmaps to Explore City/Firm Interactions in World Cities. *Comput. Environ. Urban Syst.* **2019**, *76*, 57–68. [[CrossRef](#)]
40. Chen, Q.; Jiang, R.; Yang, C.; Cai, Z.; Fan, Z.; Tsubouchi, K.; Shibasaki, R.; Song, X. DualSIN: Dual Sequential Interaction Network for Human Intentional Mobility Prediction. In Proceedings of the 28th International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2020; ACM: New York, NY, USA, 2020; pp. 283–292.
41. Suzuki, S.; Takagi, M.; Takeda, S.; Tanida, R.; Kimata, H. Deep Feature Compression with Spatio-Temporal Arranging for Collaborative Intelligence. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 3099–3103.
42. Suleman, M.A.R.; Shridevi, S. Short-Term Weather Forecasting Using Spatial Feature Attention Based LSTM Model. *IEEE Access* **2022**, *10*, 82456–82468. [[CrossRef](#)]
43. Yang, L.; Huang, R.; Huang, J.; Lin, T.; Wang, L.; Mijiti, R.; Wei, P.; Tang, C.; Shao, J.; Li, Q.; et al. Semantic Segmentation Based on Temporal Features: Learning of Temporal–Spatial Information from Time-Series SAR Images for Paddy Rice Mapping. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [[CrossRef](#)]
44. Weinberger, K.; Dasgupta, A.; Langford, J.; Smola, A.; Attenberg, J. Feature Hashing for Large Scale Multitask Learning. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; ACM: New York, NY, USA, 2009; pp. 1113–1120.
45. Jiang, R.; Song, X.; Fan, Z.; Xia, T.; Wang, Z.; Chen, Q.; Cai, Z.; Shibasaki, R. Transfer Urban Human Mobility via POI Embedding over Multiple Cities. *ACMIMS Trans. Data Sci.* **2021**, *2*, 1–26. [[CrossRef](#)]
46. Cai, P.; Yang, L.; Sun, Y. Spatio-Temporal Attention Model with Prior Knowledge for Solar Wind Speed Prediction. In *Artificial Neural Networks and Machine Learning—ICANN 2023*; Iliadis, L., Papaleonidas, A., Angelov, P., Jayne, C., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2023; Volume 14262, pp. 344–355, ISBN 978-3-031-44200-1.

47. Park, J.; Park, Y.; Kim, C.-I. TCAE: Temporal Convolutional Autoencoders for Time Series Anomaly Detection. In Proceedings of the 2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN), Barcelona, Spain, 5–8 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 421–426.
48. Zhang, Y. A Better Autoencoder for Image: Convolutional Autoencoder. Available online: https://users.cecs.anu.edu.au/~Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58.pdf (accessed on 16 October 2023).
49. Amato, F.; Guignard, F.; Robert, S.; Kanevski, M. A Novel Framework for Spatio-Temporal Prediction of Environmental Data Using Deep Learning. *Sci. Rep.* **2020**, *10*, 22243. [[CrossRef](#)]
50. Su, R.; Huang, W.; Ma, H.; Song, X.; Hu, J. SGE NET: Video Object Detection with Squeezed GRU and Information Entropy Map. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 689–693.
51. Yu, B.; Yin, H.; Zhu, Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; International Joint Conferences on Artificial Intelligence Organization: Stockholm, Sweden, 2018; pp. 3634–3640.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.