Information Fusion 53 (2020) 123–133

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Urban big data fusion based on deep learning: An overview

Jia Liu[a,b,c], Tianrui Li[a,b,c,*], Peng Xie[a,b,c], Shengdong Du[a,b,c], Fei Teng[a,b,c], Xin Yang[a,b,c]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[b] Institute of Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
[c] National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 611756, China

ARTICLE INFO

Keywords:
Urban computing
Big data
Data fusion
Deep learning

ABSTRACT

Urban big data fusion creates huge values for urban computing in solving urban problems. In recent years, various models and algorithms based on deep learning have been proposed to unlock the power of knowledge from urban big data. To clarify the methodologies of urban big data fusion based on deep learning (DL), this paper classifies them into three categories: DL-output-based fusion, DL-input-based fusion and DL-double-stage-based fusion. These methods use deep learning to learn feature representation from multi-source big data. Then each category of fusion methods is introduced and some examples are shown. The difficulties and ideas of dealing with urban big data will also be discussed.

1. Introduction

Our life and the city we live in affect each other. In the era of big data, it is urgent to effectively use urban big data to solve problems in the city, such as traffic congestion [1,2], noise pollution [3,4], air pollution [5,6], etc., to improve our life experience. Nowadays, many urban computing methods based on deep learning have been put forward to solve urban problems, such as urban traffic flow prediction [7,8], urban crowd flows prediction [9,10], urban air prediction [11,12], urban water quality prediction [13,14], etc. In these urban computing methods, the big data used by the researchers are all from different sources, such as meteorological stations, taxi detectors, online weather web sites, etc. Moreover, urban big data shows different representations, such as text, numbers and symbols. Bello et al. [15] and Zhang et al.[16] summarized five characteristics of big data, that is, large volume, large velocity, large variety, veracity and value, which are called 5V's features. The 5V's features of the data indirectly indicate a big explosion in data amount. On the one hand, how to sense, obtain and manage these big data is a challenge; On the other hand, how to analyze and excavate the value of these big data is another significant challenge. Apparently, the urban big data with 5V's characteristics brings great challenges to urban computing. Fig. 1 depicts the urban big data.

Firstly, urban big data comes from many sources. When studying the real-time city-wide traffic volume, the data usually come from taxi sensor, exploratory data, monitoring data and Internet web data. For example, Meng et al. [7] collected data from three ways, which are 155 road segments deployed with loop detectors, real-time GPS readings of 6918 taxicabs and road network and point of interest (*POI*) in Guiyang, to infer the urban traffic volume. When predicting city-wide crowd flows, we can obtain data from mobile phone signals, Internet web data, exploratory data and so on. For example, Zhang et al. [10] obtained data by two ways for predicting urban crowd flows, namely, Beijing's taxicab GPS data and meteorology data to obtain dataset TaxiBJ, and NYC bike system to obtain dataset BikeNYC. Secondly, urban big data is heterogeneous, which is reflected in different types and different existing fields. On the one hand, urban big data presents different types. Urban big data includes spatial data, temporal data, static data, dynamic data and attribute data. For example, when studying the real-time urban traffic volume, the road network belongs to spatial data, day of week belongs to temporal data, point of interest (*POI*) belongs to static data, the traffic flow of each road at different time intervals belongs to dynamic data, and the number of road lanes belongs to attribute data. On the other hand, urban big data exists in many fields. Urban big data includes social media data, traffic data, geographic data, meteorological data and other data. For example, Yao et al. [17] used traffic data, geographic data and meteorological data to predict urban taxi demand. Specific domain data are used to study specific domain issues, and the relationships between different domains cannot be ignored. Besides, urban big data is multi-modal. Different data representations, data units and data densities show the multi-modal of urban big data. For example, when studying the real-time urban air quality, data of different models is used, including text data, numerical data and so on. For example, Yi et al. [11] used three datasets to predict air quality, namely, air quality data, weather forecast data and meteorological data. Air quality data consists of the concentration of six pollutants: $PM_{2.5}$, PM_{10}, NO_2, CO, O_3 and SO_2. Weather forecast data consists of weather, temperature, wind strength and wind direction. Meteorological data consists of weather (sunny, cloudy, overcast, foggy, snow, small rain, moderate rain and heavy rain), humidity, temperature, pressure, wind speed and wind

* Corresponding author.
E-mail address: trli@swjtu.edu.cn (T. Li).

https://doi.org/10.1016/j.inffus.2019.06.016
Received 18 February 2019; Received in revised form 4 June 2019; Accepted 9 June 2019
Available online 10 June 2019

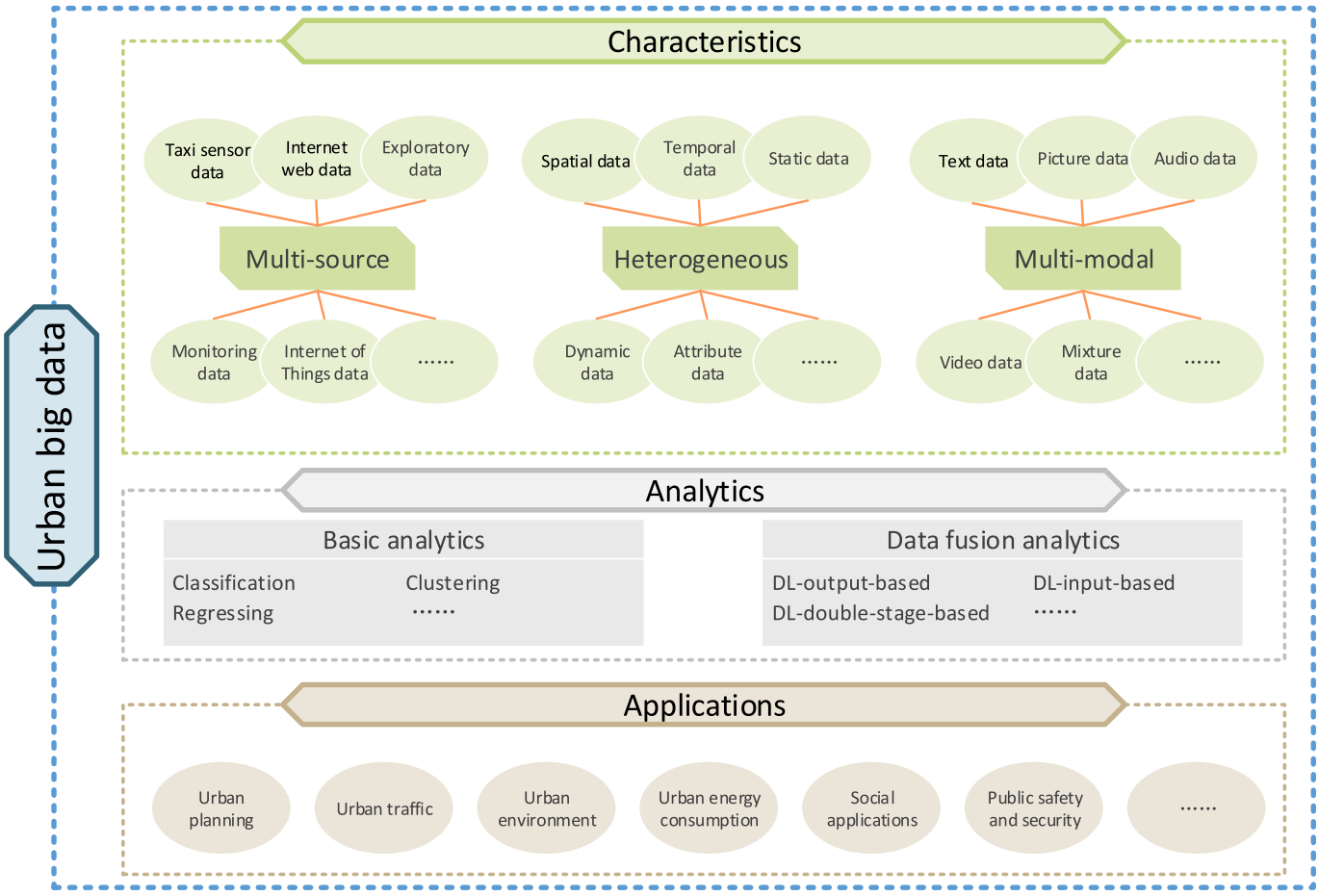


Fig. 1. Urban big data.

direction. Specific values for each indicator may be numeric, textual, or otherwise. Although the characteristics of urban big data bring specific challenges for us to analyze urban big data, urban big data will have many applications through basic analysis and fusion analysis. For example, urban big data can be used in urban planning, urban traffic, urban environment, urban energy consumption, social application and urban public safety and security [18].

To fuse the urban big data, these urban computing methods use deep learning to learn feature representation, which are found to be useful in classification and information retrieval tasks. This paper summarizes the urban big data fusion methodologies by classifying them into three categories: DL-output-based fusion, DL-input-based fusion and DL-double-stage-based fusion. The first category of data fusion methodologies trains spatial-temporal data through the deep learning model, and fuses the output of all models by feature-level-based data fusion [19], such as direct concatenation. We call them DL-output-based fusion. The second category of data fusion methodologies, which is called DL-input-based fusion, fuses data through deep fusion network while training the spatial-temporal data, and then fuses the outputs. The above two fusion models are similar in fusion process. The third category of data fusion methodologies, which is called DL-double-stage-based fusion, focuses on the stage of fusion by considering both the early fusion stage and the late fusion stage.

The contributions of our work are listed as follows:

(1) We introduce spatial-temporal data with intrinsic properties and data types, and give some instances of spatial-temporal data. Moreover, some common data fusion methods are also discussed, such as feature-based data fusion method, stage-based data fusion method and semantic meaning-based data fusion method.
(2) We summarize some existing urban big data fusion methods based on deep learning model and divide them into three categories, namely, DL-output-based fusion, DL-input-based fusion and DL-double-stage-based fusion.
(3) We briefly describe the difficulties of urban big data fusion from four aspects, namely, data quality, data sparsity, multi-modal data and spatial-temporal data, and present some new ideas of data fusion based on deep learning model.

The rest of this paper is organized as follows: in Section 2, the paper introduces spatial-temporal data and some common data fusion methods; The Section 3 shows the urban big data fusion methods based on deep learning. The difficulties and ideas of urban big data fusion are introduced in Section 4. The Section 5 is the conclusion of this paper.

2. Related work

2.1. Spatial-temporal data

A large amount of data are generated from different fields every day, which are studied by different fields, including: climate science, neuroscience, earth science, social science, physical health and transportation [20]. According to the spatial and temporal dimensions, the data can be divided into data with temporal attribute, data with spatial attribute and data with temporal and spatial attributes (spatial-temporal data), as shown in Fig. 2.

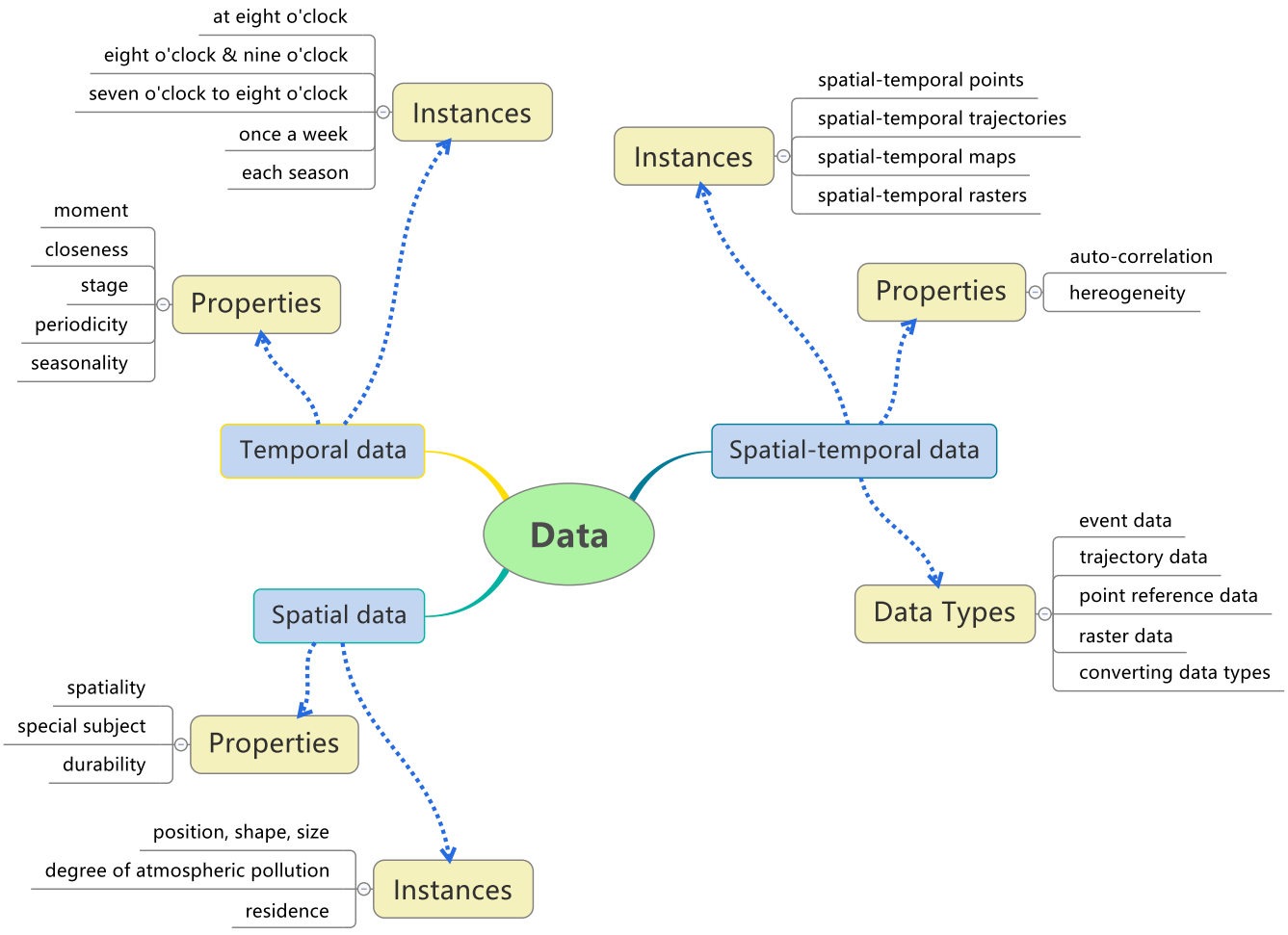


Fig. 2. Spatial-temporal data.

In general, data with temporal property is word that describes time, such as minute, second, etc. The properties of such data are moment, closeness, periodicity, seasonality and stage. For example, the data ***eight o'clock*** shows the moment. There is some relationship between data ***eight o'clock*** and data ***nine o'clock***, which shows the closeness; Data ***once a week*** shows periodicity; The data ***every season*** shows seasonality; The data ***seven o'clock to eight o'clock*** shows the stage. Data with temporal attribute is indispensable in studying the evolution and development of things.

The data with spatial property describes an object that does not change with time in a short period generally, such as road, building, etc. This kind of data has the properties of spatiality, special subject, durability and so on. Spatial property is unique to geographic information systems or spatial information systems, which refers to geometric features such as the position, shape and size, as well as the spatial relationship with adjacent things. A spatial position can be described by coordinates. Special subject attributes refers to the attribute characteristics of spatial phenomena or spatial targets other than temporal and spatial characteristics, such as the degree of atmospheric pollution. Durability property refers to an object that does not change with time in a short period, such as residence.

The development of things in different fields is essentially temporal and spatial. Therefore, most of the data are generated in the developmental process of things contain both temporal and spatial properties, which is called spatial-temporal data. Spatial-temporal data have the properties of auto-correlation and heterogeneity [20]. Auto-correlation refers to the observations made at space and time are not independent and are correlated with each other. Thus, it is crucial to account for the structure of auto-correlation among observations while analyzing spatial-temporal data. The high variability of data types and formats is the embodiment of data heterogeneity. For example, the datasets that are used to study traffic congestion, e.g., road network data, point of interest (*POI*), have different types and formats, and often come from different sources. There are various types of spatial-temporal data, including event data, trajectory data, point reference data, raster data, converting data types, etc.

Since spatial-temporal data has both time and space dimensions, it is different from data with only time dimension or space dimension. Many data mining methods widely use single dimensional data instances. However, these methods do not perform well when dealing with spatial-temporal data, as instances are structurally related in space and time, and show different attributes in spatial region and time period. Ignoring these dependencies in spatial-temporal data analysis may lead to poor accuracy of the proposed models that deal with spatial-temporal data. The amount of data with only time or space dimension is much smaller than that of spatial-temporal data, but the influence of the former on the proposed model is not lower than that of the latter. Therefore, the study of spatial-temporal data needs to be efficiently combined with spatial data and temporal data to achieve better results.

Obviously, the urban spatial-temporal big data is a part of the above data, which has all their features. The urban spatial-temporal big data includes urban temporal data, urban spatial data, event data, trajectory

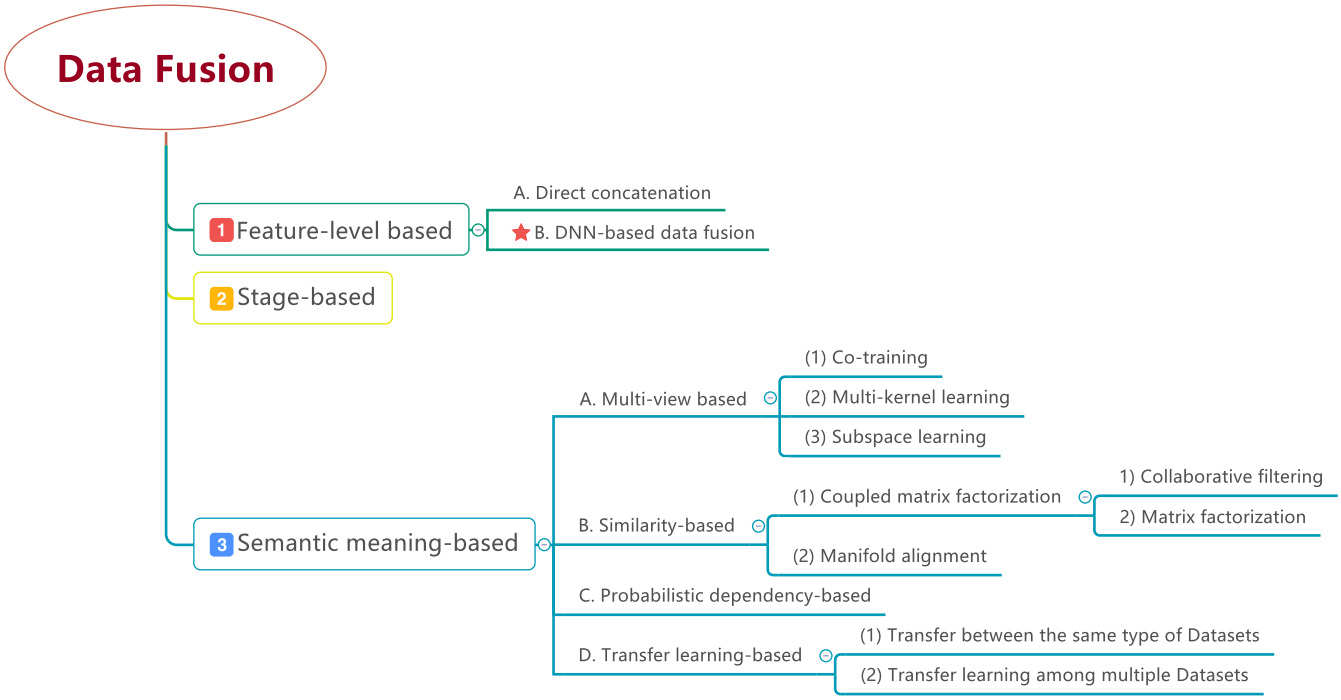


Fig. 3. Common data fusion methods [19].

data, climate data, environmental data and traffic dynamics data. The research of urban spatial-temporal big data is a basic of urban computing.

2.2. Common data fusion methods

Data fusion has a long history and is of great significance in mining data value. Early data fusion methods transformed data into a single, feature-based data, and treated the transformed data as a single dataset [21,22]. Nowadays, common data fusion methods study the value of data from different perspectives. For example, the simplest data fusion method is combining two one-dimensional datasets directly that have the same meaning in this dimension. In addition, data fusion can be carried out from the perspective of extracting the features of different dimensions. Based on the existing semantic understanding of text data, data fusion can also be conducted from the semantics of data. Different data fusion methods can bring different optimization results to machine learning model. According to different data fusion methods, Zheng divided data fusion methods into three categories [19], as shown in Fig. 3.

In feature-based data fusion methods, features of the same dimension are usually extracted from different data, and then these features are directly concatenated [23,24] or studied using deep learning methods [8,25,26]. For the method of directly combining features, several problems need to be noted: First, when directly merging data, it is necessary to remove duplicate features; Second, some features of different dimensions, which make the model have good performance, may be lost because of directly merging; Third, direct merging of features may result in overfitting. The data fusion method based on deep learning makes the deep learning model achieve good results in feature extraction and feature learning, which will be introduced in Section 3 in detail.

The stage-based data fusion method [27–29] divides the problem into different stages, then analyzes the problems of each stage through the data of this stage, and finally merges the outputs of each stage problem. For the stage-based data fusion method, the following problems need to be paid attention to: First, dividing the target problem into different stages will lead to the loss of connections between the problems in different stages; Second, how to set the roughness of the stage; Third, when combining the solutions of each stage problem, how to optimize the combination? In general, different data fusion methods usually have different effects on the combined results.

The semantic meaning-based data fusion method is studied from the semantics of data. Data contains knowledge. The similarity and correlation of knowledge contained in data and measured in different ways is the key of semantic data fusion. Data fusion method based on semantic is divided into four categories, which are multi-view based data fusion method [12,30], similarity-based data fusion method [31,32], probabilistic dependency-based data fusion method [33,34] and transfer learning-based data fusion method [35,36], respectively. The multi-view based data fusion method studies an object from the knowledge of different views. In terms of the process of researching an object, the multi-view based data fusion method is divided into co-training method [37,38], multi-kernel learning method [12,39] and subspace learning method [40,41]. Co-training method uses knowledge from different views to simultaneously train a model. The multi-kernel learning method is based on machine learning, which uses different kernels in different machine learning methods to learn. Subspace learning method learns potential subspace from different views through assuming the input views are generated from this latent subspace. The similarity-based data fusion method is usually to measure the correlation degree of multi-source data, quantify the similarity degree and construct the similarity matrix to study. Coupled matrix factorization [31,42] and manifold learning [32,43] are two classical data fusion methods based on similarity, which can find interesting structures from one dimension of data. Katz et al. [32] extracted the variable source data by two or more sensors, and then proposed a method based on manifold learning to capture the internal structure of the data, making the proposed model dependent on minimal prior knowledge. The probabilistic dependency-based data fusion method is based on graph structure. This method takes different data as nodes firstly, and then the relationship between data, such as causality, is measured by the edge (edge is divided into directed edge and undirected edge, which is determined according to the structure graph used). After constructing the structure graph of the data, some

methods of studying the structure graph are used to fuse the data. Zheng et al. [34] used spatial-temporal data from different domains to detect urban collective anomalies. They viewed a region as a document, different types of datasets as words, road network data and *POI* as keywords, and latent functions as topics, and proposed a multi-source latent-topic model to fuse data using the topic model. Transfer learning-based data fusion method mainly makes use of the concept of transfer that applies the learned knowledge to other problems. Transfer learning-based data fusion method is divided into transfer learning between the same type of datasets and transfer learning among multiple datasets. For the previous approach, the data can be transferred from a domain to another one where training data is limited. For the latter approach, the knowledge of multiple datasets can be transferred from a source to a target domain.

3. Urban big data fusion based on deep learning

There are many existing urban computing methods, and the urban data used by these methods are also diverse. Deep learning methods that have achieved good results in visual and image classification [44,45] are now also used to analyze urban big data. Compared with the models that analyse from a single dimension of a dataset, such as temporal dimension [46,47] and the models that simply fuse spatial-temporal data [48,49], the deep learning method combined with the spatial-temporal big data [8–10] is more effective in urban computing. Urban computation is defined as a process of acquisition, integration, and analysis of big and heterogeneous data generated by diverse sources in urban spaces to tackle the major issues in the city by Zheng et al. [18]. This Section elaborates some existing urban big data fusion methods based on deep learning.

3.1. Deep learning for urban big data fusion

Deep neural network (DNN) is not new in the field of artificial intelligence [19], but the application of DNN data fusion method to urban computing is a relatively new and hot research. DNN is a neural network with a large number of parameters and many hidden layers, such multi-layer perceptron (MLP). From the neural network based on back-propagation algorithm (BP algorithm), which does not work well with multiple hidden layers, to the feedback neural network, such as Restricted boltzmann machines (RBM), Convolutional neural network (CNN) and Recurrent neural network (RNN), DNN has more advantages in learning new feature representation [50], and has been proved to be superior to hand-crafted features [19,51,52]. The application of DNN in urban big data feature representation makes a good breakthrough in feature-level-based data fusion, especially in urban computing. The urban big data fusion based on deep learning has achieved many successful cases, and solved most problems in cities, including urban traffic volume prediction [7,8,53], urban crowd flows prediction [9,10,12], urban water quality prediction [13,14,54], etc. The urban big data used by these prediction models are all similar, but the methods of data fusion are different. The existing big data fusion models based on deep learning can be roughly divided into three types, which are deep learning output-based fusion (DL-output-based fusion), deep learning input-based fusion (DL-input-based fusion) and deep learning double-layer fusion (DL-double-stage-based fusion), as shown in Fig. 4. The DL-output-based fusion is used to train spatial-temporal data through the deep learning model, and then fuses the output of all models. The method for fusion is usually weighted output, and the weighted parameters are learnable parameters. The DL-input-based fusion model fuses data through deep fusion network while training the spatial-temporal data, and then fuses the outputs. The late fusion in the DL-input-based data fusion model is similar to the DL-output-based data fusion model. The DL-double-stage-based fusion model uses deep fusion network in both the early fusion stage and the late fusion stage. The similarities and differences among the three categories of data fusion methods are shown in Table 1. These three categories of data fusion methods have two fusion processes, namely early fusion and late fusion. In addition, since these three categories of data fusion methods are based on the deep learning model, they are all feature-based fusion. For urban problems, such as air pollution [55], these methods use temporal data, spatial data and external data. However, compared with DL-input-based fusion and DL-double-layer fusion, DL-output-based fusion does not use deep learning in the fusion process, but only uses other fusion methods, such as direct connection, to fuse the output of the deep learning model. Moreover, compared with DL-input-based fusion and DL-output-based fusion, the DL-double-stage-based fusion uses deep learning both in early fusion and late fusion.

3.2. The DL-output-based fusion

The DL-output-based fusion method trains spatial-temporal data through the deep learning model, and fuses the output of all models by feature-level-based data fusion [19], such as direct concatenation. Zhang et al. [10] and Yao et al. [17] used DL-output-based fusion method to predict urban crowd flows and urban taxi demand, respectively.

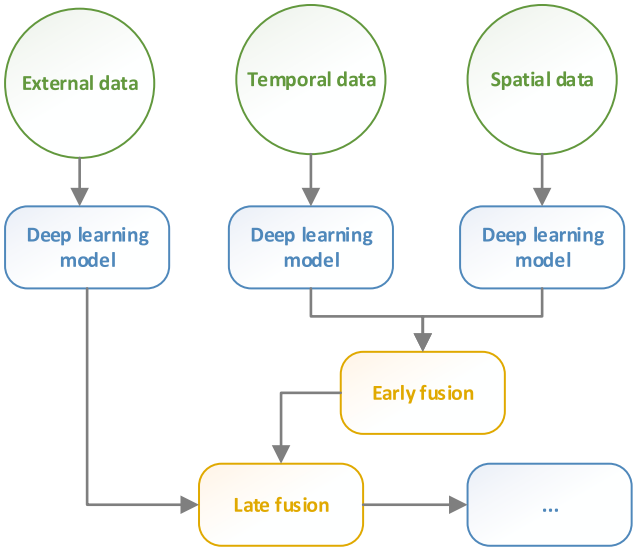
DL-output-based Fusion Case 1: Urban crowd flows prediction. Zhang et al. [10] proposed the deep spatial-temporal residual networks (ST-ResNet), which is a deep-learning-based method, to collectively predict the inflow and outflow of crowds in each region of a city. The simplified ST-ResNet is shown in Fig. 5. The spatial dependencies between any two regions in a city are modeled by the convolution-based residual networks, and the temporal properties, such as closeness, period and trend, are modeled by three residual networks, respectively. In addition, the external datasets, such as weather, holiday, event and metadata, are modeled by a two-layer fully connected neural network. After getting the output of each component, ST-ResNet uses a two-level fusion to fuse the output, namely early fusion and late fusion, as shown in Fig. 5. In early fusion, Zhang et al. [10] proposed a parametric-matrix-based method that took into account the different regions may have different degrees of influence by closeness, period and trend. The idea of feature weighting is similar to the idea of the online feature weighting mechanism in the classifier *rClass* proposed by Pratama et al.[56]. The early fusion is shown below:

$$X_{Res} = W_c \circ X_c^{L+2} + W_p \circ X_p^{L+2} + W_q \circ X_q^{L+2}, \tag{1}$$

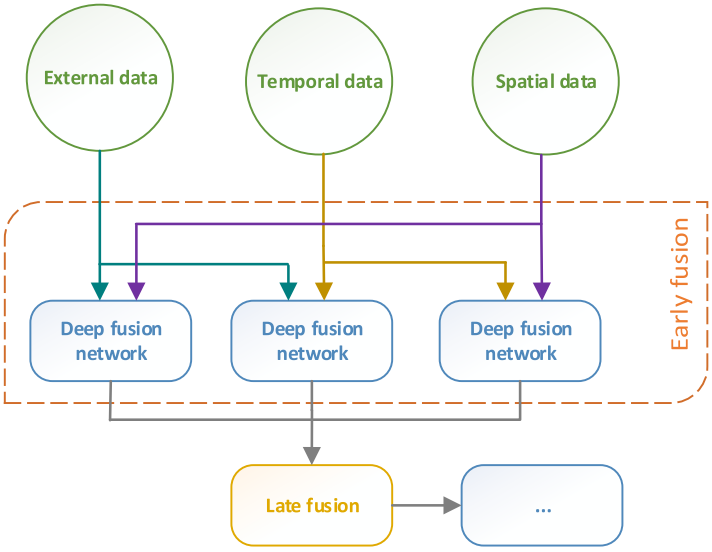
where X_{Res} is the result of early fusion, and X_c, X_p and X_q are the historical observations of temporal closeness, period and trend, respectively. L is the number of the residual units. $\circ$ is Hadamard product. W_c, W_p and W_q are the weighted parameters, which are learnable parameters. The late fusion merges the X_{Res} and the output of external component X_{Ext} directly, namely, $X_{Res} + X_{Ext}$, because X_{Ext} has been mapped to the same dimension as X_{Res} in external component.

DL-output-based Fusion Case 2: Urban taxi demand prediction. Yao et al. [17] proposed a deep multi-view spatial-temporal network (DMVST-Net), which consists of temporal view, spatial view and semantic view, to model both temporal and spatial relations. The simplified DMVST-Net is shown in Fig. 6. In spatial view, they treat one location with its surrounding neighborhood as one $S \times S$ image having one channel. The K-layer convolutional neural network is used for training, then a flatten layer is applied to transform the output to an eigenvector, and finally a fully connected layer is employed to reduce the dimensional of spatial representation. In temporal view, the region representation s_t^i and context features e_t^i are concatenated firstly, and then the result of the connection is used for an input to the Long Short–Term Memory (LSTM), which is temporal view component. In semantic view, a semantic graph of location is constructed and each node of the graph is encoded into a low dimensional vector through a graph embedding method. Then the output eigenvectors are fed into a fully connected layer. In early fusion, the spatial representations and the context features are concatenated in the temporal component, as shown in Formula (2).

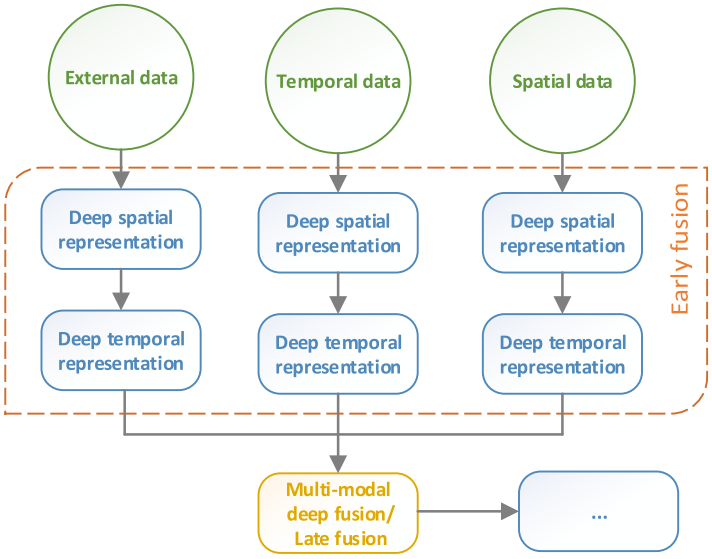
$$g_t^i = e_t^i \oplus s_t^i, \tag{2}$$



(a) The DL-output-based fusion



(b) The DL-input-based fusion



(c) The DL-double-stage-based fusion

Fig. 4. The urban big data fusion models based on deep learning.

Table 1
Similarities and differences among the three categories of data fusion methods. (Y represents Yes and N represents No).

Data fusion methods	Fusion stages	Is feature-based fusion or not	Data	Use deep models for fusing data or not	Output-based or Input-based fusion
DL-output-based	Early fusion Late fusion	Y	Temporal data Spatial data External data	N	Output-based
DL-input-based	Early fusion Late fusion	Y	Temporal data Spatial data External data	Y	Input-based
DL-double-stage-based	Early fusion Late fusion	Y	Temporal data Spatial data External data	Y	Output-based Input-based

where g_t^i is the output of concatenating the region representation s_t^i and context features e_t^i, as well as the input of LSTM. i is the number of location and t is the time interval. In late fusion, the three views are joined together by combining the output of semantic view and the output of LSTM, as shown in Formula (3). q_t^i is the result of concatenating the output of LSTM, which contains both effects of temporal and spatial view h_t^i, and the output of the semantic view m^i.

$$q_t^i = m^i \oplus h_t^i, \tag{3}$$

3.3. The DL-input-based fusion

The DL-input-based fusion method fuses data through deep fusion network while training the spatial-temporal data, and then fuses the outputs. The late fusion process in the input-based data fusion model is

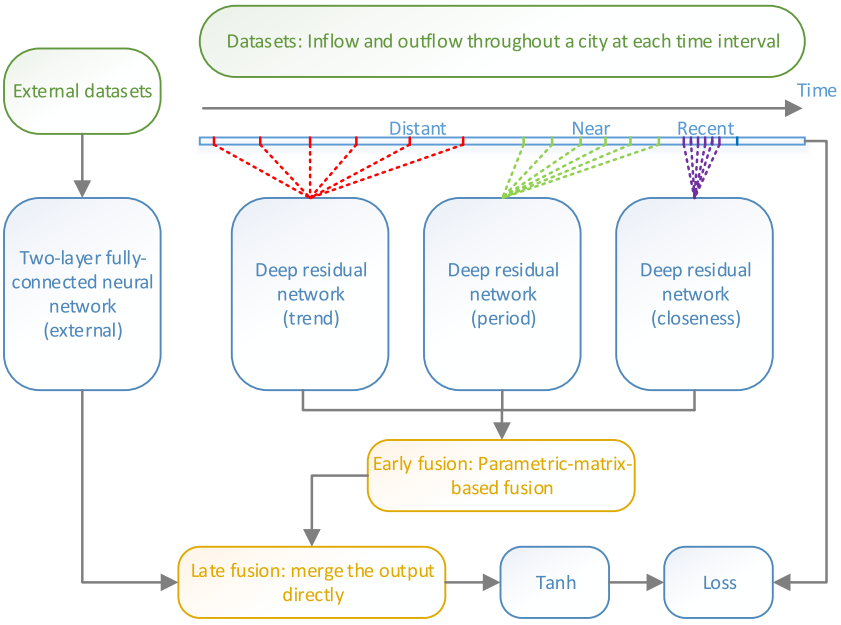


Fig. 5. Simplified ST-ResNet architecture. (The original ST-ResNet architecture is available in [10]).

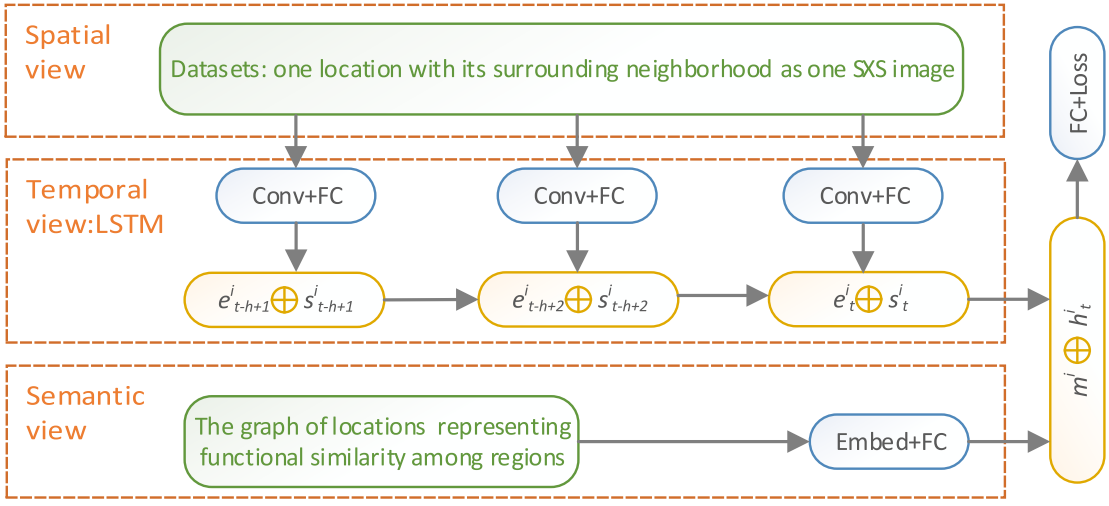


Fig. 6. Simplified DMVST-Net architecture. (The original DMVST-Net architecture is available in [17]. ⊕ presents the concatenation operate).

similar to the output-based data fusion model. Yi et al. [11] used DL-input-based fusion method to predict urban air quality.

DL-input-based Fusion Case 1: Urban air quality prediction. Yi et al. [11] proposed a deep neural network-based method (DeepAir), which consists of a deep distributed fusion network and a spatial transformation component. The simplified deep distributed fusion Network is shown in Fig. 7. The proposed spatial transformation component converts the spatial sparse air quality data into a consistent input (named AQIs) by the spatial partition, spatial aggregation and spatial interpolation because of the air pollutants' spatial correlations. In the deep distributed fusion network, there are one main feature, AQIs and four auxiliary features, meteorology, weather forecast, other pollutants, time and station ID. The main feature and four auxiliary features are fused through the FusionNet, and then the five features are fused together in a parallel manner. In early fusion, FusionNet is used to fuse multi-sources heterogeneous data. FusionNet is composed of concatenate layer, full connection layer, residual layer and full connection layer. According to different inputs, FusionNet can be considered as five sub-networks in deep distributed fusion network, which are historical weather sub-network (HW), weather forecast sub-network (WF), secondary production sub-network (SP), meta property sub-network (MP) and holistic influence sub-network (HI), and its' corresponding outputs are y_{hw}, y_{wf}, y_{sp}, y_{mp} and y_{hi}. In late fusion, considering the different effects of five sub-network outputs on the predicted results, a parametric-matrix-based fusion method, which was proposed in [10], is used to model the dynamic influences. The late fusion formula is shown below.

$$\hat{y} = Sigmoid(y_{hw} \circ w_{hw} + y_{wf} \circ w_{wf} + y_{sp} \circ w_{sp} + y_{mp} \circ w_{mp} + y_{hi} \circ w_{hi}), \quad (4)$$

where $\hat{y}$ is the predicted result and $\circ$ is Hadamard product. w_{hw}, w_{wf}, w_{sp}, w_{mp} and w_{hi} are the weights of five sub-networks and learnable parameters.

3.4. The DL-double-stage-based fusion.

The DL-double-stage-based fusion method uses deep fusion network in both the early fusion stage and the late fusion stage, which is different from DL-input-based fusion and DL-output-based fusion. Du et al. [8] and Zhang et al. [57] used DL-double-stage-based fusion method to predict urban traffic flow and urban crowd flows, respectively.

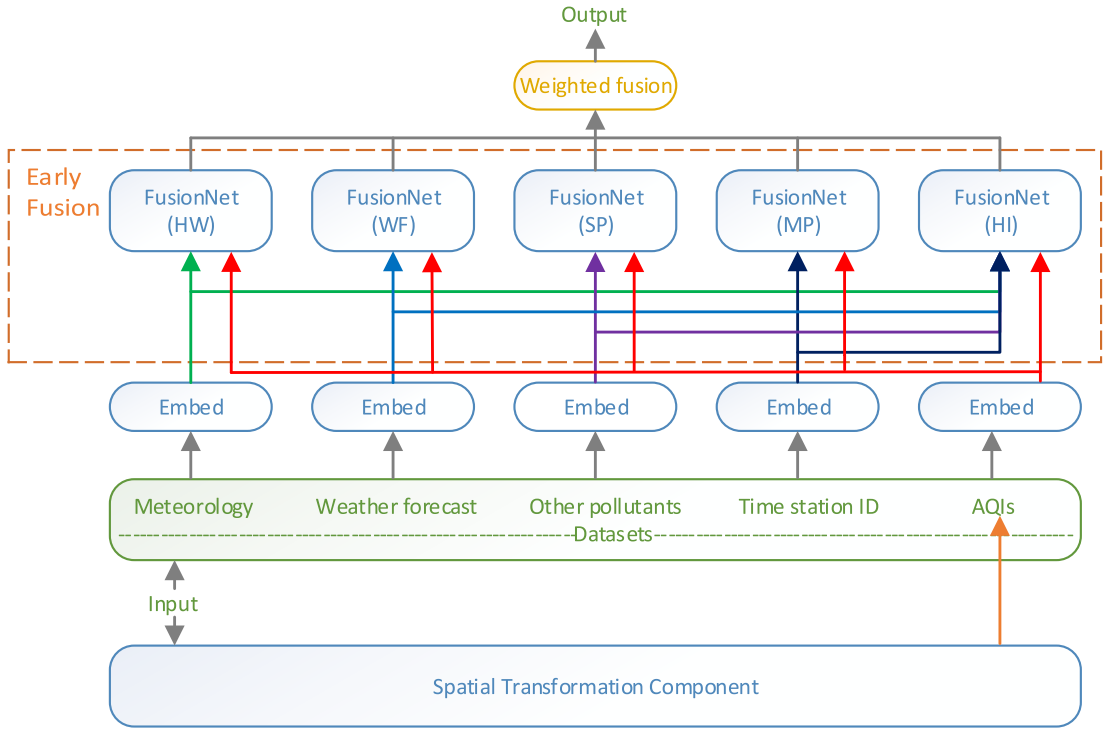


Fig. 7. Simplified deep distributed fusion network. (The original deep distributed fusion Network is available in [11]).

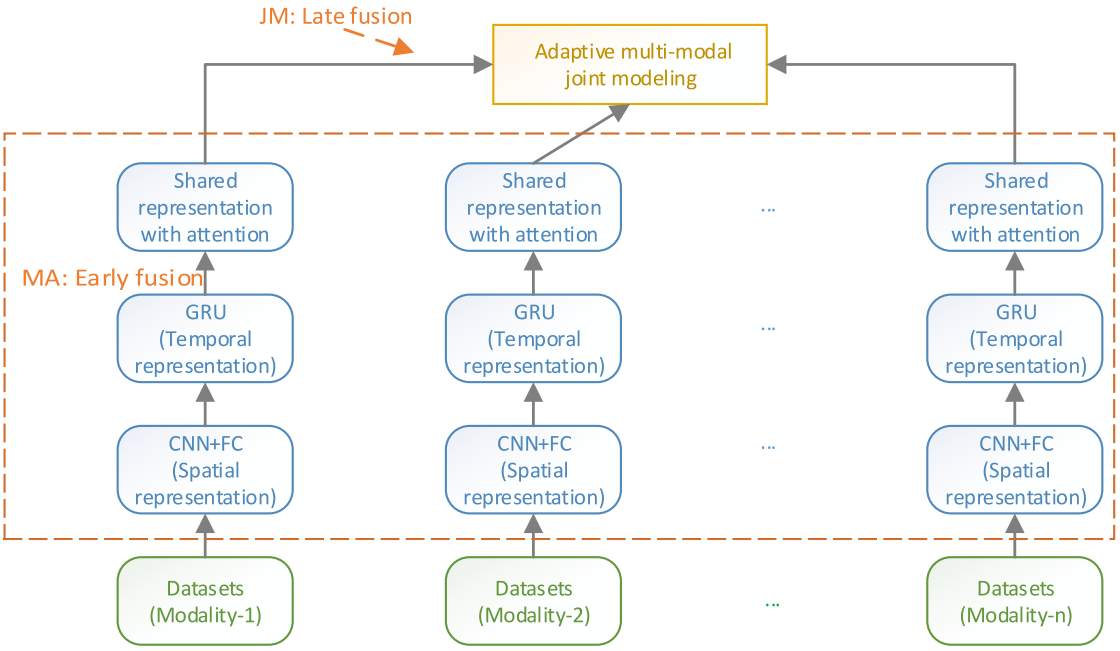


Fig. 8. Simplified hybrid multi-modal deep learning framework for traffic flow forecasting diagram. (The original hybrid multi-modal deep learning framework for traffic flow forecasting diagram is available in [8]).

DL-double-stage-based Fusion Case 1: Urban traffic flow prediction. Du et al. [8] forecasted the short-term traffic flow by proposing a hybrid multi-modal deep learning framework, which consists of convolution model, GRU model and joint model, and jointly learns the spatial-temporal correlation features and interdependence of multi-modal traffic data. The simplified hybrid multi-modal deep learning framework for traffic flow forecasting diagram is shown in Fig. 8. The convolution model is used to learn the spatial feature representation of sequence numbers' local tendency. The GRU model is used to learn the time representation of long-dependency features. The final joint model is employed to learn multi-modal data representation fusion. In early fusion, the CNN and GRU models are used to extract deep correlation features, which are spatial-temporal features. The fusion process of spatial-temporal features is shown below.

$$CNN(I_i) \rightarrow S_i\ ;\quad GRU(S_i) \rightarrow S_iT_i\ ;\quad MA(S_iT_i) \rightarrow R_i, \tag{5}$$

where S_i and T_i represent the spatial and temporal correlation features, which will be obtained from each dataset I_i with CNN and GRU models,

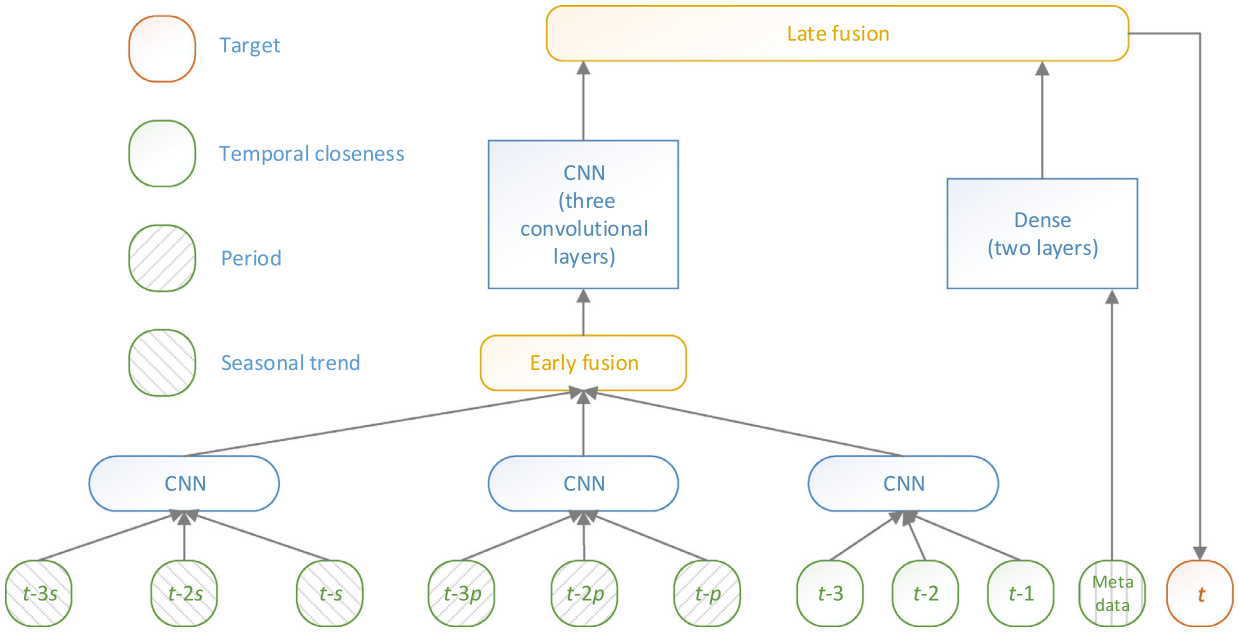


Fig. 9. Simplified DeepST architecture. (The original DeepST architecture is available in [57]).

respectively. *MA* is a multi-modal feature level fusion layers with attention mechanism and R_i is shared representation of the spatial correlation feature S_i and the temporal correlation feature T_i with attention assisted learning. In late fusion, a joint and adaptive deep learning framework is proposed to fuse spatial-temporal shared features. The multi-modal joint model is shown in Formula (6). *JM* is an adaptive multi-modal joint model, and π is the joint fusion representation for different learned spatial-temporal pair R_i. i represents each modality input. W^i and b^i are weights and biases that will be learned in the joint model.

$$JM((R_1, R_2, \dots, R_n),\ W^i,\ b^i) \rightarrow \pi,\quad i = 1, 2, \dots, n. \tag{6}$$

DL-double-stage-based Fusion Case 2: UrbanFlow system. Zhang et al. [57] built a real-time crowd flows forecasting system *UrbanFlow* by a DeepST architecture, which is composed of three components: temporal dependent instances, convolutional neural networks, and early and late fusions. The simplified DeepST architecture is shown in Fig. 9. In the first stage of DeepST, the input is generated from all temporal properties, such as temporal closeness, period and seasonality trend. In the second stage of DeepST, the CNN module is leveraged to capture spatial closeness dependency. In the final stage, early and late fusions are used to fuse different types of ST data. In early fusion, a similar domains' data is fused by a convolution layer to capture closeness, periodic and seasonality trend patterns together. The early fusion based convolution is shown in Formula (7).

$$H^{(2)} = f(W_c^{(2)} * H_c^{(1)} + W_p^{(2)} * H_p^{(1)} + W_s^{(2)} * H_s^{(1)} + b^{(2)}), \tag{7}$$

where $H^{(2)}$ is the result of fusion, and $H_c^{(1)}$, $H_p^{(1)}$, $H_s^{(1)}$ are the outputs of the first convolutional layer over closeness, periodic, trend sequences, respectively. $W_c^{(2)}$, $W_p^{(2)}$, $W_s^{(2)}$ and $b^{(2)}$ are the parameters in the second layer. In late fusion, different domains' data, such as external factors (i.e. day of week, weekday/weekend), are fused, and the external factors are considered to be the global features. The process of late fusion is shown in Formula (8). G_t is the global feature vector, and the $\hat{X}_t$ is the predicted tensor. *tanh* is a hyperbolic tangent, which squashes real numbers to range between [−1,1].

$$\hat{X}_t = tanh(W^{(5)} \cdot H^{(4)} + W_G^{(5)} \cdot G_t + b^{(5)}). \tag{8}$$

4. Difficulties and ideas of urban big data fusion

Urban big data fusion is rapidly developing with the help of deep learning, especially in urban computing. However, urban big data is very complex, and we only extract a small part of its knowledge [32,58]. Therefore, we will encounter many challenges to unlock the power of knowledge from urban big data as much as possible. This Section will give some difficulties of urban big data fusion firstly, and then some ideas of urban big data fusion will be discussed.

4.1. Difficulties of urban big data fusion

Urban big data is more abundant, but also more random, disorderly, difficult to predict and uncontrollable. Urban big data has 5V' characteristics, as described in Section 1. It can be structured data, e.g., lists, unstructured data, e.g., video, and semi-structured data, e.g., posts related to cities. The difference between data brought by the multi-source and heterogeneous nature of urban big data is a challenge for data fusion. Many existing methods [8,27,59] have been proposed and proved to be effective in the integration of urban big data. However, these methods only aim at a specific problem or specific urban big data, and only a small part of the information in the data is integrated. Therefore, more researches are needed to fuse urban big data. The difficulties of urban big data fusion are as follows.

Data quality. Data with good quality plays an important role in data fusion. The low quality of data is represented as incorrect, missing, wrong format, incomplete, and so on [60]. Take data missing as an example, there are many reasons for data missing, such as omission, unavailability, don't-care value, etc., and data missing will increase the estimation error and reduce the performance of data analysis. On the one hand, it is difficult to represent features of missing data for data fusion. On the other hand, although the missing value is filled by the missing value filling method with good performance, there is still some error that is difficult to estimate.

Data sparse. Due to many unpredictable reasons, e.g., increasing the dimension of the data, most of datasets are sparse, it is often difficult to estimate the correct distribution from limited observations. Sparse data refers to the data with most values missing or zero in the dataset, which is not useless data, but incomplete information. Generally, the larger the data size is, the more sparse the data is, while the deep learning model

needs large scale data to train its parameters. What is more, some algorithms, such as Least Squares Support Vector Machine (LSSVM), are required to use all training data for each prediction. Although some algorithms have been proposed to solve this problem, such as the adaptive pruning algorithm [61], the data sparse problem is still not well solved. Therefore, how to deal with the problem of big data sparsity is a very challenging task.

Multi-modal data. Different data representations, different data units and different data densities show the multi-modal of urban big data, as described in Section 1. Datasets in different fields usually have different distributions and ranges, because data from different domains consists of multiple forms with different representations, descriptions, scales, and densities. In the real world, the obtained multi-modal data has various forms and different structures. In addition, multi-modal data is often unstructured and has the characteristics of high-dimensional or even ultra-high-dimensional. The feature representation of ultra-high-dimensional data is challenging for data fusion method. How to fuse multi-modal data with high-dimensional features and extract or select the most effective features for the current task is a problem worthy of further study.

Spatial-temporal data. The nature of static correlation and dynamic evolution of data is that the process of the real world is spatial-temporal. In general, spatial-temporal data has independent spatial properties, e.g., geographical hierarchy and distance, and temporal properties, e.g., sequential, periodic and seasonality trend patterns [57]. How to represent the temporal variation trend and spatial distribution law of big data plays a great role in the spatial-temporal data fusion. Moreover, there are many methods to integrate spatial-temporal data, but it is time-consuming to find a specific method that has good performance for spatial-temporal data fusion. Improving the efficiency of big data analysis is another significant problem.

4.2. Ideas of urban big data fusion

Although urban big data fusion has many difficulties, the research on urban big data is crucial and urgent. Aiming at the difficulties in urban big data fusion, we can start from the nature of big data difficulties and solve the problems by making up for the deficiencies of big data. For example, Yi et al. [62] filled in missing values by building a spatial-temporal multiview-based learning (ST-MVL) method, which made up for the problem of data missing. We can consider the fusion of urban big data from the following directions.

To integrate data by learning the feature representation of multi-modal big data. There are some models based on deep neural network to learn the feature representation of multi-modal big data, and they have good performance in classification and information extraction. For example, a deep auto-encoder architecture was proposed to learn features over multiple modalities by Ngiam et al. [26]. On the basis of these work, we can integrate data by analyzing the feature representation of multi-modal big data at different granularities. Different granularities information that corresponding to different types of data is constructed to analyze the internal structure and relationship of multi-modal big data. To fuse multi-modal big data, we can employ deep learning to explore the deep feature representation and multi-layer feature representation of multi-modal big data. What is more, the correlation relation and sharing representation mechanism between different granularities levels also can be used to integrate data.

Although this paper classifies the urban big data fusion methods based on deep learning into DL-output-based fusion, DL-input-based fusion and DL-double-stage-based fusion, there are only two layers of data fusion in these methods, namely, early fusion and late fusion. In the actual deep learning model, there are many submodules whose input and output need to be fused. Therefore, compared with the DL-output-based fusion, DL-input-based fusion and DL-double-stage-based fusion methods, the DL-multi-stage-based fusion method will be considered to study the fusion of urban big data with different granularities. In addition, a hybrid DL-multi-stage-based fusion method can be considered in deep learning model by combining other fusion methods and DL-multi-stage-based fusion method in a special case.

Finally, on the basis of the correlation relation and shared representation of features among different granularities of urban big data, we consider to fuse the data through model fusion, that is, model combination or cross-model combination (model can be CNN, RNN, LSTM, RBM, etc.) to realize the deep learning of multi-task integration, and complete the data fusion at the same time. For example, we can make the spatial and temporal attributes of the data fuse well by combinating models the CNN and RNN.

5. Conclusion

In this paper, we attempt to provide a broad overview of the urban big data fusion based on deep learning, which is hot and challenging. Urban big data from different aspects are analyzed firstly, and spatial-temporal data and some common data fusion methods that can be roughly divided into three categories are briefly described. Then, some existing multi-modal urban big data fusion methods based on deep learning are classified into three categories, which are DL-output-based fusion, DL-input-based fusion and DL-double-stage-based fusion, and described separately. Finally, according to the attributes and characteristics of urban big data, the difficulties and some ideas in studying urban big data are provided. Apparently, there are lots of literatures on urban big data fusion methods based on deep learning, and we only cover part of the work in this rapidly developing field. We still hope that this article will be helpful to future research.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 61773324).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2019.06.016.

References

[1] M.R. Jabbarpour, H. Zarrabi, R.H. Khokhar, S. Shamshirband, K.-K.R. Choo, Applications of computational intelligence in vehicle traffic congestion problem: a survey, Soft Comput. 22 (7) (2018) 2299–2320.
[2] R. Arnott, T. Rave, R. Schöb, Alleviating Urban Traffic Congestion, 1, MIT Press Books, 2005.
[3] J. Bello, C. Silva, O. Nov, R.L. Dubois, A. Arora, J. Salamon, C. Mydlarz, H. Doraiswamy, Sonyc: a system for monitoring, analyzing, and mitigating urban noise pollution, Commun. ACM 62 (2) (2019) 68–77.
[4] Y. Xu, Y. Zhu, Z. Qin, Urban noise mapping with a crowd sensing system, Wirel. Netw. (2018) 1–14.
[5] N. Gendron-Carrier, M. Gonzalez-Navarro, S. Polloni, M.A. Turner, Subways and urban air pollution, Technical Report, National Bureau of Economic Research, 2018.
[6] S. Yu, P. Li, L. Wang, Y. Wu, S. Wang, K. Liu, T. Zhu, Y. Zhang, M. Hu, L. Zeng, et al., Mitigation of Severe Urban Haze Pollution by a Precision Air Pollution Control Approach, 8, Scientific Reports (Nature Publisher Group), 2018, pp. 1–11.
[7] C. Meng, X. Yi, L. Su, J. Gao, Y. Zheng, City-wide traffic volume inference with loop detector data and taxi trajectories, in: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2017, p. 1.
[8] S. Du, T. Li, X. Gong, Z. Yu, S.-J. Horng, A hybrid method for traffic flow forecasting using multimodal deep learning, arXiv preprint arXiv:1803.02099 (2018).
[9] W. Jin, Y. Lin, Z. Wu, H. Wan, Spatio-temporal recurrent convolutional networks for citywide short-term crowd flows prediction, in: Proceedings of the 2nd International Conference on Compute and Data Analysis, ACM, 2018, pp. 28–35.
[10] J. Zhang, Y. Zheng, D. Qi, Deep spatio-temporal residual networks for citywide crowd flows prediction., in: Proceedings of the 31th AAAI Conference on Artificial Intelligence, 2017, pp. 1655–1661.
[11] X. Yi, J. Zhang, Z. Wang, T. Li, Y. Zheng, Deep distributed fusion network for air quality prediction, in: Proceedings of the 24th SIGKDD Conference on Knowledge Discovery and Data Mining, 2018, pp. 965–973.
[12] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, T. Li, Forecasting fine-grained air quality based on big data, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 2267–2276.

[13] H. Assem, S. Ghariba, G. Makrai, P. Johnston, L. Gill, F. Pilla, Urban water flow and water level prediction based on deep learning, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 317–329.

[14] Y. Liu, Y. Zheng, Y. Liang, S. Liu, D.S. Rosenblum, Urban water quality prediction based on multi-task multi-view learning, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, AAAI Press, 2016, pp. 2576–2582.

[15] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: recent achievements and new challenges, Inf. Fusion 28 (2016) 45–59.

[16] L. Zhang, Y. Xie, L. Xidao, X. Zhang, Multi-source heterogeneous data fusion, in: Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data, IEEE, 2018, pp. 47–51.

[17] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, Z. Li, J. Ye, D. Chuxing, Deep multi-view spatial-temporal network for taxi demand prediction, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI press, 2018, pp. 2588–2595.

[18] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban computing:concepts, methodologies, and applications, ACM Trans. Intell. Syst. Technol. 5 (3) (2014) 1–55.

[19] Y. Zheng, Methodologies for cross-domain data fusion: an overview, IEEE Trans. Big Data 1 (1) (2015) 16–34.

[20] G. Atluri, A. Karpatne, V. Kumar, Spatio-temporal data mining: a survey of problems and methods, ACM Comput. Surv. (CSUR) 51 (4) (2018) 83.

[21] P. Pavlidis, J. Weston, J. Cai, W.S. Noble, Learning gene functional classifications from multiple data types., J. Comput. Biol. 9 (2) (2002) 401–411.

[22] P. Maragos, P. Gros, A. Katsamanis, G. Papandreou, Cross-modal integration for performance improving in multimedia: a review, in: Proceedings of the IEEE International Conference on Image Processing, 2008, pp. 3412–3416.

[23] Y. Fu, Y. Ge, Y. Zheng, Z. Yao, Y. Liu, H. Xiong, J. Yuan, Sparse real estate ranking with online user reviews and offline moving behaviors, in: Proceedings of the 2014 IEEE International Conference on Data Mining, IEEE, 2014, pp. 120–129.

[24] Z. Wang, D. Zhang, X. Zhou, D. Yang, Z. Yu, Z. Yu, Discovering and profiling overlapping communities in location-based social networks., IEEE Trans. Syst. Man Cybern. 44 (4) (2014) 499–509.

[25] M. Pratama, J. Lu, S. Anavatti, E. Lughofer, C.-P. Lim, An incremental meta-cognitive-based scaffolding fuzzy neural network, Neurocomputing 171 (2016) 89–105.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 689–696.

[27] L. Zhu, F. Guo, J.W. Polak, R. Krishnan, Urban link travel time estimation using traffic states-based data fusion, IET Intell. Transp. Syst. 12 (7) (2018) 651–663.

[28] Y. Zheng, Y. Liu, J. Yuan, X. Xie, Urban computing with taxicabs, in: Proceedings of the 13th international conference on Ubiquitous computing, ACM, 2011, pp. 89–98.

[29] B. Pan, Y. Zheng, D. Wilkie, C. Shahabi, Crowd sensing of traffic anomalies based on human mobility and social media, in: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2013, pp. 344–353.

[30] A. Kumar, H. Daumé, A co-training approach for multi-view spectral clustering, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 393–400.

[31] J. Shang, Y. Zheng, W. Tong, E. Chang, Y. Yu, Inferring gas consumption and pollution emission of vehicles throughout a city, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014, pp. 1027–1036.

[32] O. Katz, R. Talmon, Y.-L. Lo, H.-T. Wu, Alternating diffusion maps for multimodal data fusion, Inf. Fusion 45 (2018) 346–360.

[33] N.J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, H. Xiong, Discovering urban functional zones using latent activity trajectories, IEEE Trans. Knowl. Data Eng. 27 (3) (2015) 712–725.

[34] Y. Zheng, H. Zhang, Y. Yu, Detecting collective anomalies from multiple spatio-temporal datasets across different domains, in: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2015, p. 2.

[35] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, Y. Yu, Translated learning: Transfer learning across different feature spaces, in: Advances in Neural Information Processing Systems, 2009, pp. 353–360.

[36] P. Yang, W. Gao, Multi-view discriminant transfer learning., in: International Joint Conference on Artificial Intelligence, 2013, pp. 1848–1854.

[37] Y. Zheng, X. Chen, Q. Jin, Y. Chen, X. Qu, X. Liu, E. Chang, W.-Y. Ma, Y. Rui, W. Sun, A cloud-based knowledge discovery system for monitoring fine-grained air quality, preparation, Microsoft Tech Report, 2014. http://research.microsoft.com/apps/pubs/default.aspx.

[38] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Computational Learning Theory, ACM, 1998, pp. 92–100.

[39] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, J. Mach. Learn. Res. 12 (7) (2011) 2211–2268.

[40] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2085–2098.

[41] N. Chen, J. Zhu, E.P. Xing, Predictive subspace learning for multi-view data: a large margin approach, in: Proceedings of the Advances in Neural Information Processing Systems, 2010, pp. 361–369.

[42] V.W. Zheng, Y. Zheng, X. Xie, Q. Yang, Collaborative location and activity recommendations with gps history data, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 1029–1038.

[43] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, E. Chang, Diagnosing new york city's noises with ubiquitous data, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, 2014, pp. 715–725.

[44] Z. Zhong, J. Li, Z. Luo, M. Chapman, Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework, IEEE Trans. Geosci. Remote Sens. 56 (2) (2018) 847–858.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.

[46] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, L. Damas, Predicting taxi–passenger demand using streaming data, IEEE Trans. Intell. Transp. Syst. 14 (3) (2013) 1393–1402.

[47] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, Z. Wang, Prediction of urban human mobility using large-scale taxi traces and its applications, Front. Comput. Sci. 6 (1) (2012) 111–121.

[48] Y. Tong, Y. Chen, Z. Zhou, L. Chen, J. Wang, Q. Yang, J. Ye, W. Lv, The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 1653–1662.

[49] D. Deng, C. Shahabi, U. Demiryurek, L. Zhu, R. Yu, Y. Liu, Latent space model for road networks to predict time-varying traffic, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1525–1534.

[50] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[51] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[52] U. Brefeld, T. Scheffer, Co-em support vector learning, in: Proceedings of the 21th International Conference on Machine Learning, ACM, 2004, p. 16.

[53] Y.-J. Lee, O. Min, Long short-term memory recurrent neural network for urban traffic prediction: a case study of seoul, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2018, pp. 1279–1284.

[54] Y. Liu, Y. Liang, S. Liu, D.S. Rosenblum, Y. Zheng, Predicting urban water quality with ubiquitous data, arXiv preprint arXiv:1610.09462 (2016).

[55] M. Picornell, T. Ruiz, R. Borge, P. García-Albertos, D. de la Paz, J. Lumbreras, Population dynamics based on mobile phone data to improve air pollution exposure assessments, J. Exposure Sci. Environ. Epidemiol. 29 (2) (2019) 278.

[56] M. Pratama, S.G. Anavatti, J. Lu, Recurrent classifier based on an incremental metacognitive-based scaffolding algorithm, IEEE Trans. Fuzzy Syst. 23 (6) (2015) 2048–2066.

[57] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, Dnn-based prediction model for spatio-temporal data, in: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2016, p. 92.

[58] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, A. Rocha, Multimodal data fusion for sensitive scene localization, Inf. Fusion 45 (2019) 307–323.

[59] M.R. Shahrbabaki, A.A. Safavi, M. Papageorgiou, I. Papamichail, A data fusion approach for real-time traffic state estimation in urban signalized links, Transp. Res. Part C 92 (2018) 525–548.

[60] J.P. Gouveia, J. Seixas, G. Giannakidis, Smart city energy planning: Integrating data and tools, in: Proceedings of the 25th International Conference Companion on World Wide Web, 2016, pp. 345–350.

[61] X. Yang, J. Lu, G. Zhang, Adaptive pruning algorithm for least squares support vector machine classifier, Soft Comput. 14 (7) (2010) 667–680.

[62] X. Yi, Y. Zheng, J. Zhang, T. Li, St-mvl: filling missing values in geo-sensory time series data, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, 2016, pp. 2704–2710.