ISPRS Journal of Photogrammetry and Remote Sensing 218 (2024) 19–32

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Multi-level urban street representation with street-view imagery and hybrid semantic graph

Yan Zhang [a,b], Yong Li [b,c], Fan Zhang [b,*]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, 430079, China
[b] Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing, 100871, China
[c] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region of China

ARTICLE INFO

Keywords:
GeoAI
Street-view imagery
Urban scene representation
Multi-modal
Spatial interaction
Traffic estimation

ABSTRACT

Street-view imagery has been densely covering cities. They provide a close-up perspective of the urban physical environment, allowing a comprehensive perception and understanding of cities. There has been a significant amount of effort to represent the urban physical environment based on street view imagery, and this representation has been utilized to study the relationships between the physical environment, human dynamics, and socioeconomic environments. However, there are two key challenges in representing the urban physical environment of streets based on street-view images for downstream tasks. First, current research mainly focuses on the proportions of visual elements within the scene, neglecting the spatial adjacency between them. Second, the spatial dependency and spatial interaction between streets have not been adequately accounted for. These limitations hinder the effective representation and understanding of urban streets. To address these challenges, we propose a dynamic graph representation framework based on dual spatial semantics. At the intra-street level, we consider the spatial adjacency relationships of visual elements. Our method dynamically parses visual elements within the scene, achieving context-specific representations. At the inter-street level, we construct two spatial weight matrices by integrating the spatial dependency and the spatial interaction relationships. It could account for the hybrid spatial relationships between streets comprehensively, enhancing the model's ability to represent human dynamics and socioeconomic status. Furthermore, aside from these two modules, we also provide a spatial interpretability analysis tool for downstream tasks. A case study of our research framework shows that our method improves vehicle speed and flow estimation by 2.4% and 6.4%, respectively. This not only indicates that street-view imagery provides rich information about urban transportation but also offers a more accurate and reliable data-driven framework for urban studies. The code is available at: (https://github.com/yemanzhongting/HybridGraph).

1. Introduction

Urban streets, carrying a significant amount of human activity, are essential units for urban analysis, allowing for more precise capture of human activities and uncovering hidden knowledge within urban areas (Zhu et al., 2017; Zhang et al., 2024b). Quantifying urban streets, specifically building the computational representation of streets, is crucial to understanding the coupling, interactions, and long-term evolution between the urban physical environment and human dynamics and socio-economic environment (Biljecki and Ito, 2021; Fan et al., 2023). There has been extensive work attempting to represent the urban physical environment based on street view imagery and utilize this representation to study the relationships between the physical environment and urban design (Salazar-Miranda et al., 2023; Zhang et al., 2023), human behaviors (Zhang et al., 2019), public health (Kang et al., 2020), human perception (Zhang et al., 2018b), crime (Zhang et al., 2021a), and other factors.

However, existing research that employs street-view imagery for the representation of urban streets has notable limitations. First, these studies do not consider the spatial relationships between the visual elements within the street. When representing the urban environment based on street-view imagery, current research often focuses only on the proportion or number of individual visual elements (such as the proportion of trees) in isolation, without considering how these elements relate to each other in terms of their spatial arrangement or adjacency within the street scene. In other words, they do not account for the spatial context in which these elements occur (Zhang et al., 2018a). This simplistic representation fails to fully capture the complexity of the urban physical environment when addressing specific urban issues.

* Corresponding author.
E-mail addresses: sggzhang@whu.edu.cn (Y. Zhang), yong.li@connect.ust.hk (Y. Li), fanzhanggis@pku.edu.cn (F. Zhang).

https://doi.org/10.1016/j.isprsjprs.2024.09.032
Received 16 May 2024; Received in revised form 4 August 2024; Accepted 26 September 2024
Available online 18 October 2024

Indeed, spatial relationships between elements within a street, such as their relative positions and layouts, have a significant impact on how the city street is perceived and functions. For example, if two street-view images have the same proportion of trees, but in one image the trees are evenly scattered while in the other image, the trees are clustered at one end, this will result in significantly different shading, aesthetics, and potential pedestrian flow in these two scenes (Zhang et al., 2023a). Therefore, the first research gap is how to dynamically model the spatial adjacency of visual elements in the images of a street in order to incorporate this information into downstream tasks. We refer to this process as intra-street level representation.

Second, when representing streets at the inter-street level, there is a lack of adequate consideration for the spatial dependency and interaction between streets (Liu and Biljecki, 2022). On the one hand, adjacent streets often exhibit similarities in appearance, human activity, and socioeconomic attributes, which should be considered when representing streets. This similarity is based on the first law of geography, which states that everything is related to everything else, but near things are more related than distant things. This principle is referred to as spatial dependency. On the other hand, even streets that are far apart can have strong connections in terms of human spatial interaction, such as between workplaces and residential areas, or between groups of people who are connected in social networks and whose physical interactions correspond to long-distance connections between two streets (Zheng et al., 2023). However, when estimating human activity or socioeconomic status based on streets, there is not enough understanding of the spatial relationships between streets and how they influence each other. This lack of understanding can lead to inadequate estimates and an incomplete representation of the complex relationships between streets. Addressing these gaps emphasizes incorporating spatial effects and geographic rules into AI models. GeoAI (Geospatial Artificial Intelligence) has been proved could enhance the accuracy and robustness of related tasks. Therefore, the second research gap of this study lies in quantifying these two spatial relationships between streets and fusing them to represent spatial context more accurately.

To address these challenges, this paper proposes a dynamic urban graph representation framework based on dual spatial semantics. At the intra-street level, we designed a dynamic weighting module that incorporates the spatial adjacency of visual elements. At the inter-street level, we establish two graphs: one representing the spatial dependencies in real-world space, and the other capturing the long-distance spatial interactions in the space of human activities. We then fuse these graphs to represent the human dynamics and socioeconomic status.

Specifically, at the intra-street level, addressing the issue of neglecting spatial adjacency relationships between visual elements in existing research, we first compute semantic segmentation proportions with clear physical meanings in street-view imagery. Subsequently, we construct a normalized matrix based on the pixel adjacency relationships of elements and weight the semantic features accordingly. The weighted vectors not only include the visual proportions of various elements in the scene but also their spatial positions and distributions. Furthermore, considering the varying importance of different elements in different tasks (e.g., the proportion of trees is more important in health-related tasks), we introduce an $n \times n$ attention map matrix for dynamic adjustment (assuming the dimension of visual elements is n). This ultimately achieves an adaptive effect of "same elements, different semantics”.

At the inter-street level, we fuse dual spatial semantics graphs to address the neglect of spatial context between streets in current research. The first layer of spatial semantics considers the first- or second-order neighbors of street nodes, using k-nearest neighbors or Queen adjacency to construct a spatial weight matrix, thereby capturing this spatial dependency. The second layer of spatial semantics is more intricate, extending beyond traditional neighborhood effects and having a closer relationship with human activities (Askarizad et al., 2021). It outlines potential connections between different spatial units while considering the strength of these connections. This layer of spatial semantics is based on users' mobility trajectories in social media to construct a spatial weight matrix. Based on these two spatial weight matrices, we designed two feature maps to extract spatial semantic information, followed by feature fusion using a concatenation-based method. Ultimately, we propose a dynamic urban street representation method based on dual spatial semantics. It dynamically analyzes visual elements while considering the mixed spatial relationships between nodes, beyond the simple logic of “stronger influence with spatial proximity”.

Therefore, we propose an integrated framework that includes two levels of spatial thinking addressing the aforementioned research gaps, as well as an interpretability analysis tool that corresponds to the semantic and spatial interpretability aspects of the two modules, respectively. Our proposed framework can be applied to various downstream tasks related to human dynamics and socioeconomic status, such as transportation (Hu et al., 2023), epidemics, land use, urban vitality, and geographical health (Zhang et al., 2021b; Han et al., 2023).

To validate our proposed dynamic urban graph representation framework, we take traffic estimation as a case study, which is highly related to human dynamics and socioeconomic status. For example, streets adjacent to main roads often experience significantly increased traffic flow, exemplifying a typical spatial dependence effect. Furthermore, as urbanization progresses, increased commuting times and job-housing imbalances affect road traffic conditions not only in the immediate vicinity but also in more distant areas (Pan and Lai, 2019; Meng et al., 2023; Zhang et al., 2022c). This commuting pattern between residential and work areas represents a typical form of spatial interaction (Askarizad et al., 2021), playing a crucial role in urban systems. We explored the potential influence by using traffic flow volume estimation and vehicle speed estimation as downstream tasks and investigated the interpretable effects of different visual elements on traffic flow estimation. The results show that our method outperforms traditional models by 2.4% and 6.4% in traffic speed and volume estimation (Zhang et al., 2022a), respectively, validating the effectiveness of our approach. The model's spatial interpretability analysis indicates that the traditional model significantly underestimates traffic congestion levels in high-density urban centers when spatial interaction efforts are not considered.

As shown in Fig. 1, this work contributes to the field of urban visual intelligence and GeoAI by providing a framework for street representation considering multi-level spatial effects, which can be applied to a wide range of urban-related downstream tasks (Wu and Biljecki, 2023).

2. Background

The evolution of street-view imagery has been profoundly influenced by the burgeoning map service industry and the concurrent advances in computer vision. These factors jointly foster their application within urban understanding and perception studies. We discuss three relative research backgrounds:

2.1. Spatial organization of street-view graph elements at the intra-street level

In recent years, the demand for higher-level spatial understanding of streets has increased with the advancements in deep learning. It moves beyond simply detecting and recognizing the classes and proportions of elements in images. This approach is known in the field of computer science as scene graph computation (Chang et al., 2021). Allows for a clear representation of elements, their attributes, and the relationships between elements within a scene. The core of spatial organization in street-view imagery lies in structuring various elements within the

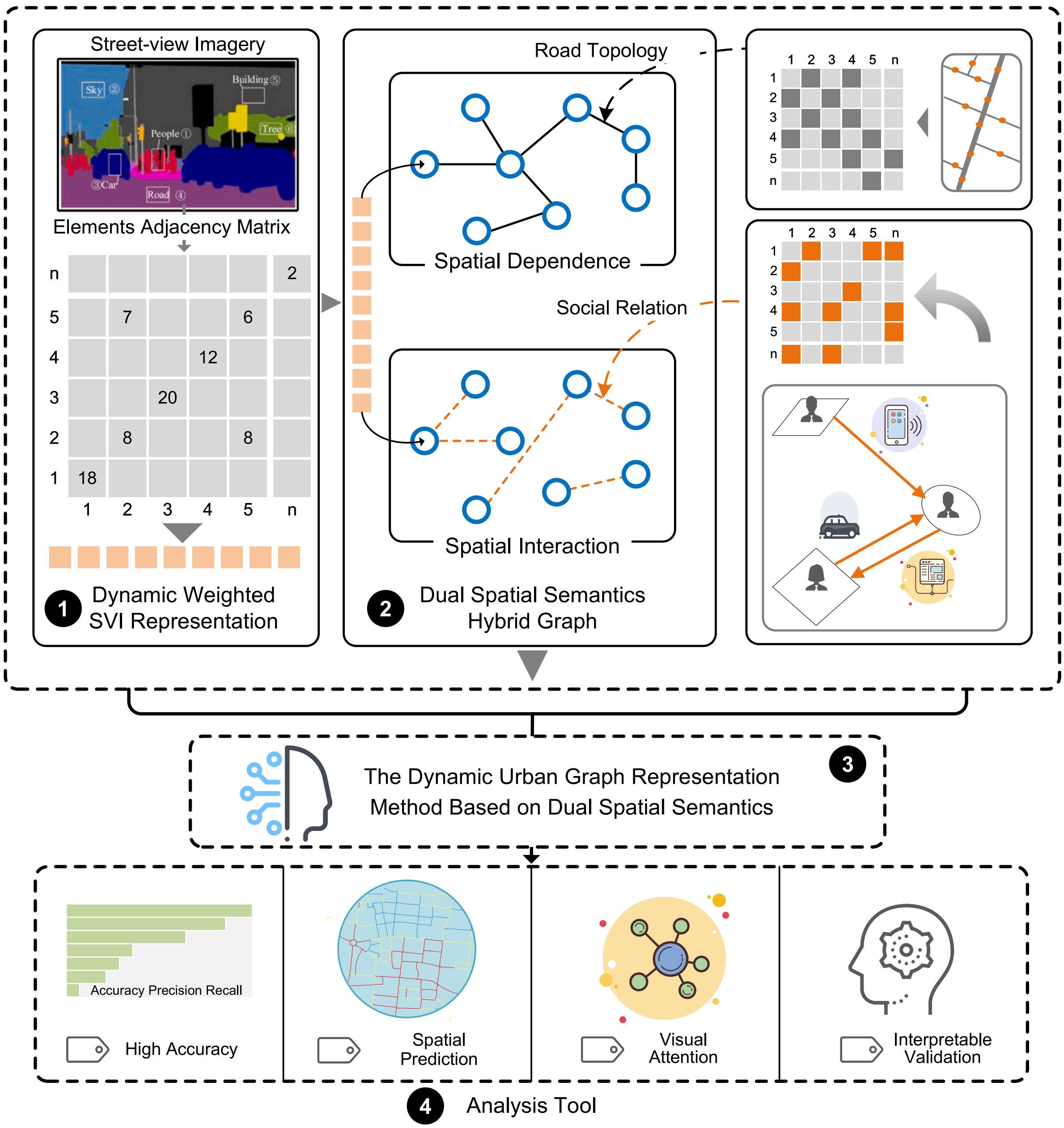

Fig. 1. The framework consists of four main components: (1) At the intra-street level, the framework computes the weighted street-view imagery semantic features based on the element adjacency matrix. (2) At the inter-street level, the framework constructs the dual spatial semantics hybrid graph, incorporating spatial dependence from the road network and spatial interaction from social relationships. (3) The framework develops the dynamic urban street graph representation method based on dual spatial semantics. (4) It provides downstream task analysis tools, including accuracy evaluation, spatial estimation, attention visualization, and interpretable analysis.

street environment. Neural networks are applied to comprehend these elements and their relationships, thereby learning to represent the elements more efficiently (Pagani et al., 2021; Grekousis, 2019). Zhang et al. (2022b) introduced a feature extraction approach that considers the spatial context. This method deconstructs street-view imagery into text-based forms using multimodal techniques, enabling the joint capture of entity-attribute relationships of spatial elements. While this approach effectively extracts spatial dependency relationships, it suffers from imprecision in describing spatial relationships, often resorting to vague terms like "near" or "beside". Another perspective involves extracting different image features, such as shallow, mid-level, and deep features, and incorporating street-view imagery depth information into the analysis (Ma et al., 2023). Through neural networks' robust fitting capability, relationships between different feature layers are constructed. However, the pixel-level granularity of features may not fully represent an image's overall characteristics, potentially leading to an incomplete representation of street-view imagery. Currently, there is a lack of methods that efficiently extract image entities and organize spatial relationships, highlighting the need to develop such methodologies to advance the analysis and application of street-view imagery research.

2.2. Integration and application of spatial semantic elements at the inter-street level

Spatial semantic information, represented by the topology of the road network, can fully incorporate spatial positional relationships between street-view imagery (Liu and Song, 2024). This information is a beneficial complement to visual image data (Xu et al., 2022b). For instance, integrating semantic information from points of interest enables

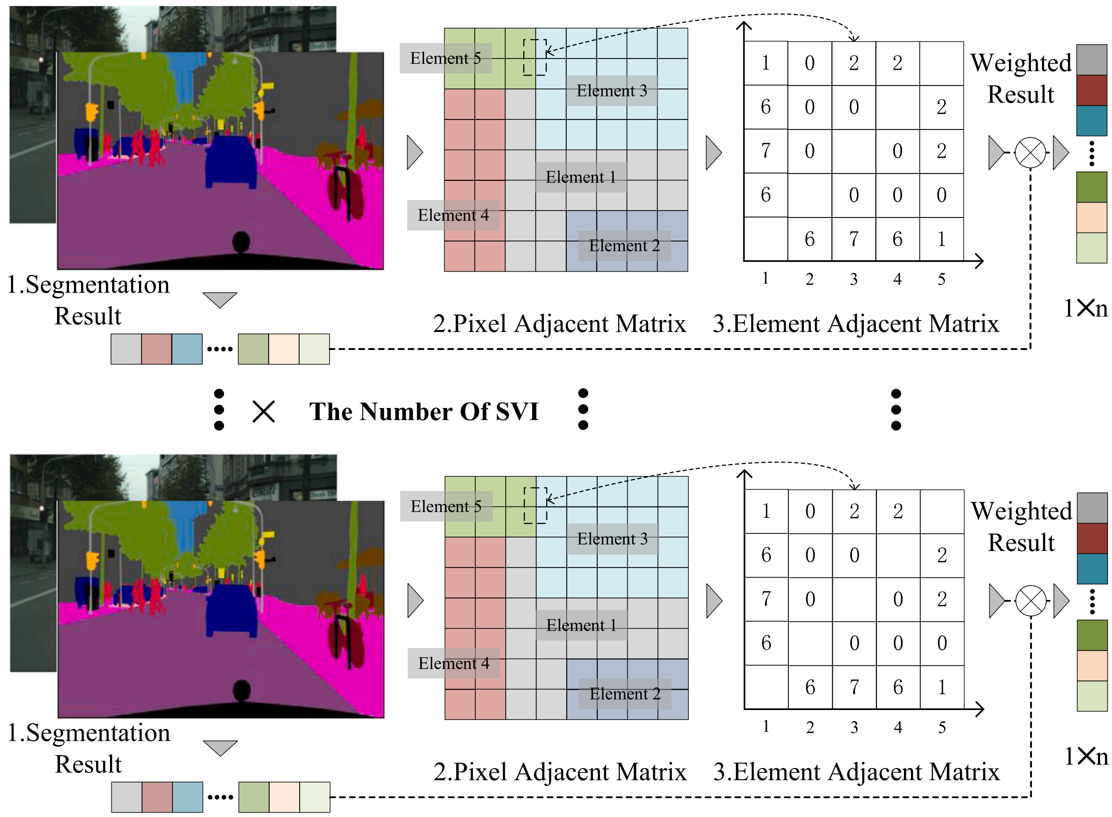

Fig. 2. Workflow diagram illustrating the process of obtaining weighted semantic features: (1) Original panoramic images are segmented into n semantic classes; (2) Each class is then pixelized to calculate element-wise distributions; (3) The proportion of each visual element is computed, resulting in a $1 \times n$ dimensional feature vector; (4) resulting the weighted semantic features.

a more comprehensive perception and exploration of the urban physical environment (Hankey et al., 2021; Xu et al., 2022a). Furthermore, given that street-view imagery in cities is continuously distributed along streets, some studies have leveraged this continuity. Focusing on the information overlapping between consecutive street-view imagery (Xu et al., 2022a), it could classify urban streets efficiently through multi-perspective and temporal integration. On a broader spatial scale, the spatial semantic information implied in urban internal migration activities, such as the commute flows (Chen et al., 2024), is also considered valuable semantic data. This spatial semantic information can be integrated with visual information from street-view imagery for a multi-scale urban perception. This insight also serves as an inspiration for current research, prompting us to consider broadening our observational scope beyond just street-view imagery and network data. Integrating human activities into our research framework, specifically through the analysis of social media data (Cao et al., 2020; Shao et al., 2021) and mobile signaling information (Li et al., 2021)—complements the visual and semantic data of urban landscapes. By integrating these diverse data sources, we could generate high-quality representations of urban regions (Huang et al., 2023) and better understand urban dynamics (Zhou et al., 2024).

2.3. Related work in estimating human dynamics and social socioeconomic status

Street-view imagery intuitively maps the urban physical environment and captures urban dynamics over a wide area at a relatively low cost (Biljecki and Ito, 2021). Although street-view imagery lacks the dynamic nature of video or trajectory data, they can still reflect traffic dynamics to some extent (Zhang et al., 2019). This is determined by two factors: first, street-view imagery effectively identifies buildings, especially residential buildings, which have a key impact on the traffic flow and volume (Pan et al., 2020). Second, street-view imagery indicates road widths, which influence drivers' route choices, as they often prefer wider main roads over narrower, busier paths—features that street-view imagery can also measure (Guan et al., 2023). These insights form the theoretical basis for our research.

Traditional traffic flow estimation methods primarily utilize trajectory data (Wang et al., 2023b,a). They rely on historical trajectory patterns to model traffic conditions, focusing on predicting future states based on past data. Graph neural networks (GNNs) incorporate the spatial topological relationships of road networks into the computation, transforming street topology into complex network structures (Yin et al., 2021). This allows for the implementation of classification or regression tasks for urban spatial analysis using GNNs (Kaczmarek et al., 2023). Integrating spatially explicit information significantly improves traffic flow predictions (De Sabbata et al., 2023; Zhang et al., 2024a). However, these approaches primarily consider adjacency relationships within road networks, addressing spatial dependencies, but not spatial interactions (Zhang et al., 2023b; Shi et al., 2024; Wang and Zhu, 2024). Long-distance spatial interactions, generated through the movement and exchange of people, goods, and information among different geographic units, must also be included in the traffic modeling (Yu and Wang, 2023).

3. Method

This section introduces the dynamic urban street graph representation method based on dual spatial semantics. It begins by constructing the weighted street-view imagery semantic features based on the element adjacency matrix within the scene. Subsequently, it forms the dual spatial semantics Hybrid Graph, which incorporates spatial dependence from the road network and spatial interaction from social relationships

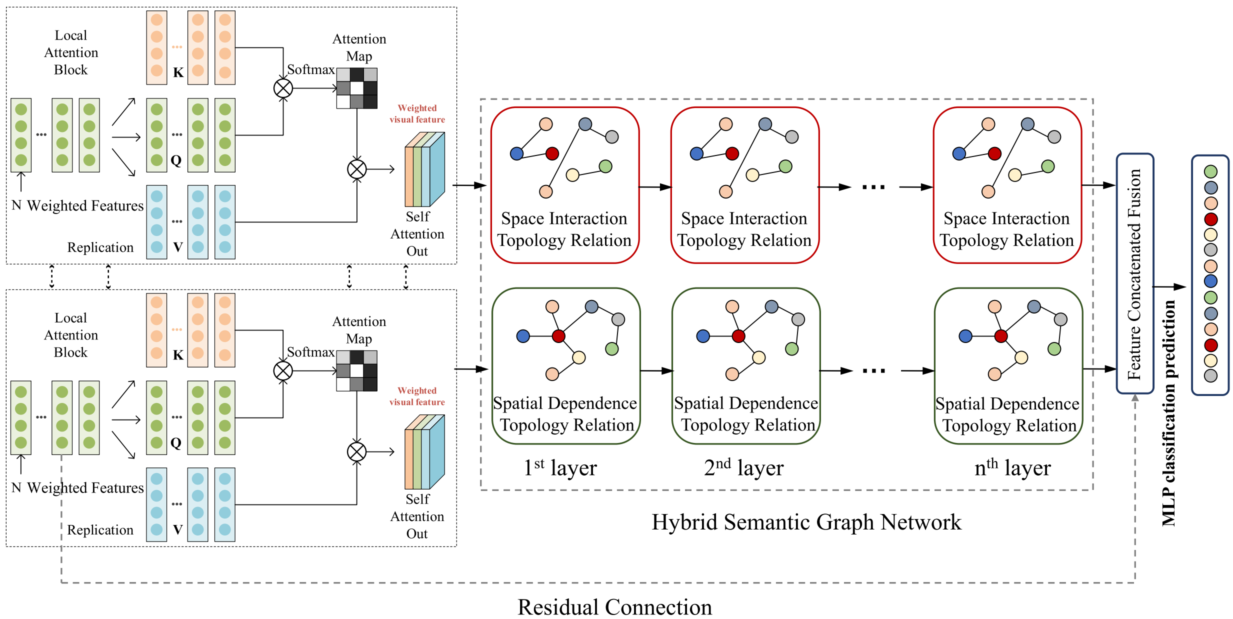

Fig. 3. The architectural schematic of the dynamic urban representation method based on dual spatial semantics. Employing a parameter-sharing strategy at shallow layers, the methodology integrates self-attention and graph convolutional layers for dynamic weighting. This approach effectively amalgamates long-distance spatial interactions with close-range spatial dependencies, augmenting the model's ability to capture both global and local traffic patterns. The incorporation of residual connections enhances model robustness. The resulting feature fusion streamlines MLP-based classification, presenting a comprehensive approach for effective learning and representation for urban street.

at the street level. Following this, the hybrid semantic graph network is used to generate multi-task street representations. Finally, the section includes the evaluation metrics for the downstream tasks.

3.1. Dynamic visual representation method at the intra-street level

In contrast to traditional semantic segmentation methods that only compute the pixel ratios of various visual elements to represent images, our approach includes adjacency relationships among visual elements. We initially obtain the initial feature values through semantic segmentation, calculating the proportions of various visual elements to form a semantic feature vector $(1 \times n)$, where n represents the number of categories of visual elements (Fig. 2, Step 1). Next, we compute the adjacency relationships at the pixel level between these visual elements. For instance, if there are two adjacent pixels between Element 3 and Element 5 in the image, then in the adjacency matrix, $A_{3,5} = 2$. We traverse all n categories of visual elements to obtain the complete adjacency matrix (Fig. 2, Step 3). The interaction strength between corresponding categories is represented by each adjacency matrix value, sized $n \times n$.

Then we perform a dot product of the original semantic feature vector with this adjacency matrix to obtain a new weighted feature vector of size $1 \times n$. This step incorporates spatial adjacency relationships, enhancing our understanding and representation of the streetscape. The weighted feature calculation formula is the following:

$$E' = E \times (I + A) \tag{1}$$

Here, E represents the original feature vector, A stands for the adjacency matrix, and I is the identity matrix. Thus, E' becomes the new weighted feature vector including both the visual and adjacency information. Utilizing this weighted visual representation vector helps to effectively overcome the limitations of traditional methods, such as when dealing with scenarios where trees are unevenly distributed or clustered at one end of the scene. It should be noted that while the features E' are more informative at this stage (considering the spatial adjacent relationship), our tests have shown that directly using E' in downstream tasks does not yield satisfactory results. Ideal performance is only achieved after fine-tuning with a small amount of task-specific data.

This feature extraction framework is adaptable, enabling users to choose different models based on specific application needs. For example, users may opt for other types of neural networks, such as deep convolutional networks, fully convolutional networks, or customized neural network architectures, depending on the desired resolution, processing speed, or amount of available training data, to optimize performance and accuracy.

At last, to adapt better to various downstream tasks, we introduced an attention mechanism to process the feature vector E'. We designed a dynamic attention module to capture the street-view imagery representation at the inter-street level (Fig. 3 left). This computation allows the model to focus on key features while ignoring irrelevant information, thus adapting to various downstream tasks (for instance, the most crucial visual elements for traffic tasks are vehicles and pedestrians, etc.). The final output of this module is the integrated image feature, weighted by an adjacency matrix and self-attention mechanisms.

Specifically, we assign physical significance to each dimension of the weighted feature vector V'. V' is a d-dimensional integrated scene feature vector comprising x_i, weighted by the scene level adjacency matrix A. For instance, x_1 might represent the physical characteristics of a tree. Subsequently, we utilize a self-attention module to dynamically adjust the importance of various physical feature factors. This module adaptively weighs each feature, customizing the image's representation to fit the specific scenario. During the computation process, the size of the embedding matrix for the input feature is $n \times d$. Within the self-attention module, the computation involves transforming x_i into the query Q_i, key K_i, and value V_i, as follows:

$$\begin{aligned} Q_i &= x_i W^Q, \\ K_i &= x_i W^K, \\ V_i &= x_i W^V, \end{aligned} \tag{2}$$

where W^Q, W^K, and W^V are learnable weight matrices that project x_i into query, key, and value spaces, respectively.

The attention scores, which dictate the weighting of each node's features, are calculated using a softmax function applied to the dot-product of queries and keys:

$$\alpha_{ij} = \frac{\exp(\mathrm{ReLU}(Q_i \cdot K_j^T))}{\sum_{k \in \mathcal{N}(i)} \exp(\mathrm{ReLU}(Q_i \cdot K_k^T))}, \tag{3}$$

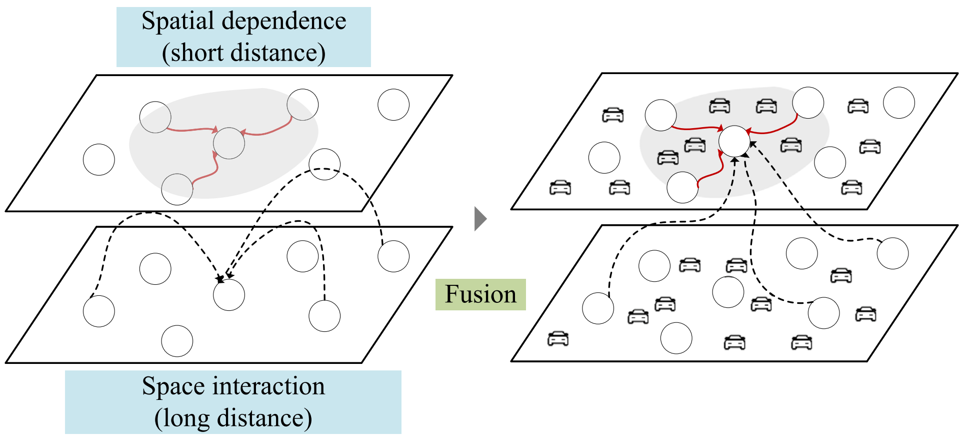

Fig. 4. Schematic representation of the integration of both spatial dependence at short distances and spatial interaction at long distances.

where α_{ij} denotes the normalized attention coefficient between nodes i and j. The weights α_{ij} form the attention map $\tilde{A}$, representing the relative importance or relevance between the query and all keys.

This module focuses on specific visual elements of the input data that are most relevant to the given task (Fan et al., 2024). Attention weights act as filters that selectively prioritize certain parts of the input, allowing the model to concentrate on the most informative features. This process effectively enhances the model's predictive capabilities by highlighting salient environmental attributes.

$$\text{Attention}(Q_i, K_i, V_i) = \sum_j \alpha_{ij} V_j, \tag{4}$$

where α_{ij} is the normalized attention score between node i and node j, and V_j is the value vector associated with node j. The attention mechanism computes a weighted sum of the value vectors V_j using the attention weights α_{ij}.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5}$$

Here, the score function typically involves a dot product and a scaling factor $\frac{1}{\sqrt{d_k}}$, where d_k is the dimensionality of the keys. The attention weights α_{ij} are derived from this score function. The dynamic scene feature values computed by the self-attention mechanism serve as the initial street representation values for the next step of graph computation.

3.2. The hybrid semantic graph network at the inter-street level

Multiple scenes of a street constitute the smallest analytical unit at the inter-street level. To capture spatial dependencies and spatial interactions simultaneously, we designed a hybrid graph neural network framework comprising two heterogeneous graphs (as shown in Fig. 3 right). The first graph targets spatial interactions (long-distance interactions), while the second graph focuses on spatial dependencies (short-distance dependencies). These two graphs jointly process the output from the previous module, and the fused features are subsequently input into a scene classification layer. The integration of both intra-street and inter-street information constitutes the overall output of our model. Here is a detailed introduction to these modules:

At the street level, as illustrated on the right side of Fig. 3, the model employs graph neural networks to perform multi-layer iterations and updates on the nodes. Users can select different encoders as needed, such as Graph Attention Networks (GAT), Graph Convolutional Network (GCN), and GraphSAGE, among others. The following is the method for updating the node features using GAT::

$$H_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)} W^{(l)} H_j^{(l)}\right) \tag{6}$$

Here, $\alpha_{ij}^{(l)}$ denote the attention coefficients and form the attention map matrix $\tilde{A}$. Each visual element in this matrix represents the attention weight between the corresponding scene nodes. The attention map capturing short-distance spatial relationships is denoted as $\tilde{A}_1$, while the attention map capturing long-distance spatial relationships is denoted as $\tilde{A}_2$. $\alpha_{ij}^{(l)}$ are computed as:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(a^{(l)T}\left[W^{(l)}H_i^{(l)} \parallel W^{(l)}H_j^{(l)}\right]\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\text{LeakyReLU}\left(a^{(l)T}\left[W^{(l)}H_i^{(l)} \parallel W^{(l)}H_k^{(l)}\right]\right)\right)} \tag{7}$$

In this equation, $\parallel$ denotes concatenation, $a^{(l)}$ is a learnable weight vector for layer l, $W^{(l)}$ is the weight matrix for layer l, σ is a non-linear activation function such as ReLU, and $H_i^{(l)}$ represents the feature vector of node i at layer l. $\mathcal{N}(i)$ denotes the neighborhood of node i. The street level matrix $\tilde{W}$ plays a critical role in this process by defining the neighborhood relationships between scenes. Only the nodes j for which $\tilde{W}_{ij} \neq 0$ are included in $\mathcal{N}(i)$.

In this paper, we need to build two spatial matrices $\tilde{W}$ to capture two spatial relationships (As shown in Fig. 4). The first matrix $\tilde{W}_1$ represents whether the road nodes are directly connected to each other. If node i is adjacent to node j, the value of w_{ij} is 1, otherwise it is 0.

$$w_{ij}^1 = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The second matrix $\tilde{W}_2$ symbolizes the long-distance relationship. If there is a flow from node i to node j, the value of w_{ij} will be equal to the number of flows; if there is no flow, we fill it with 0.

$$w_{ij}^2 = \begin{cases} m & \text{if there are } m \text{ flows from node } i \text{ to node } j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$\tilde{W}_1 = (w_{ij}^1)_{n\times n} = \begin{bmatrix} w_{11}^1 & w_{12}^1 & \cdots & w_{1n}^1 \\ w_{21}^1 & w_{22}^1 & \cdots & w_{2n}^1 \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1}^1 & w_{n2}^1 & \cdots & w_{nn}^1 \end{bmatrix} \tag{10}$$

$$\tilde{W}_2 = (w_{ij}^2)_{n\times n} = \begin{bmatrix} w_{11}^2 & w_{12}^2 & \cdots & w_{1n}^2 \\ w_{21}^2 & w_{22}^2 & \cdots & w_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1}^2 & w_{n2}^2 & \cdots & w_{nn}^2 \end{bmatrix} \tag{11}$$

In both matrices, i and j represent different scene nodes, and n stands for the number of flows from node i to node j. These weight matrices will be used to analyze the structure and dynamics of the traffic network. The **street level adjacency matrix** ($\tilde{W}_1$) reflects the basic connectivity properties of the network, while the **street level flow matrix** ($\tilde{W}_2$) provides information on the intensity of interactions

Table 1
Selected taxi trip data.

StreetID	TaxiID	TripID	Date	Lon	Lat	Direction	Speed (km/h)
2704	49 152	111	2017-07-03 05:33:51	114.210533	30.562820	282	20
2705	49 152	111	2017-07-03 05:34:17	114.207982	30.563584	295	39
7671	49 152	111	2017-07-03 05:34:47	114.205067	30.564996	351	30
7671	49 152	111	2017-07-03 05:35:17	114.206249	30.567611	21	15
7671	49 152	111	2017-07-03 05:34:51	114.205147	30.565320	15	35
…	…	…	…	…	…	…	…

Note: Direction is the angle of rotation clockwise starting from the north.

between nodes. Through these matrices, we can gain a deeper understanding of the spatial dependencies and spatial interactions within the network.

Combining the convolutional features of the graph constructed from two different spatial relationships ($H_{sd}^{(l)}$) and ($H_{si}^{(l)}$) is a crucial step in improving model performance.

$$x_{\text{cat}} = \text{Concatenate}(H_{sd}^{(l)}, H_{si}^{(l)}, H^{(0)}) \tag{12}$$

Here, $H^{(0)}$ is the initial node feature's embedding. x_{cat} effectively combines both spatial information to provide a comprehensive representation.

$$x_{\text{final}} = \text{MLP}(x_{\text{cat}}) \tag{13}$$

The concatenated feature x_{cat} is subsequently passed through a multi-layer perceptron (MLP) to obtain the final classification output x_{final}. The MLP serves as a classifier, mapping high-level features to predicted classes. By utilizing this intricate combination of self-attention, graph convolutional processes with residual learning, and multiscale feature fusion, our proposed method enhances the robustness and adaptability to various urban physical environments.

Compared to traditional methods, neural networks inherently excel at processing multi-modal and heterogeneous data. Multi-modal data encompass information captured in various forms like images, text, and sound, whereas heterogeneous data include diverse formats and sources. The information fusion method used in this paper takes into account multi-source information, including weighted visual scene features $H^{(0)}$, the spatial dependency graph features at the street $H_{sd}^{(l)}$, and spatial interaction graph features $H_{si}^{(l)}$. Neural networks are adept at integrating and interpreting these complex data streams.

3.3. Evaluation metric of street traffic state estimation downstream task

In this paper, we adopt street traffic state estimation as the downstream test task to explore the effectiveness of the proposed method. Traffic conditions directly correspond to the urban physical environment and spatial patterns, which can effectively evaluate the effectiveness of our method. Additionally, we simplify the task of traffic flow estimation to a classification task, which has a clear definition and is easy to evaluate. This also simplifies the impact of time-series data such as traffic flow to a certain extent, enabling the comprehensive testing of our model's performance.

We infer the traffic flow status based on the scene features. Streets are labeled with different categories of traffic flow status based on various classification criteria. In addition, we use three common classification metrics: Accuracy, Precision, and Recall, to quantify the model's performance.

Accuracy is a fundamental metric that measures the overall correctness of identifying different traffic conditions. It is defined as the ratio of correctly classified instances to the total number of instances:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} C_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}} \tag{14}$$

where C_{ij} is the count of instances belonging to class i and predicted as class j.

Precision focuses on the accuracy of positive predictions. It is the ratio of true positives to the total number of instances predicted to be positive for each class. Precision helps us to understand the reliability of positive predictions:

$$\text{Precision} = \frac{1}{n}\sum_{i=1}^{n} \frac{C_{ii}}{\sum_{j=1}^{n} C_{ji}} \tag{15}$$

Recall measures the ability of a classification system to capture all instances of each class. It is the ratio of true positives to the total number of actual instances for each class. Recall provides insights into the system's sensitivity to instances of each class:

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n} \frac{C_{ii}}{\sum_{j=1}^{n} C_{ij}} \tag{16}$$

In the following sections, we assess different model performances based on those metrics for the evaluation of street traffic state identification using street-view imagery in an $n - class$ setting.

4. Experiments and results

4.1. Data preparation and preprocessing

Inverting the traffic operation state based on the urban physical environment is a challenging problem, as the traffic state is affected by a variety of factors and presents a variety of manifestations. We acquired 5,115,512 cab trajectory data in the study area in July 2017, as shown in Table 1. Key information such as the vehicle's latitude, longitude, speed, direction, and order number with the street where it is located are recorded at one-minute intervals. Among them, we mainly focus on location information, as well as speed information. In addition, we collected 14,115 panoramic images, which basically cover the main roads of the study area. Semantic segmentation was performed on these panoramic images. We used the Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2016) model pre-trained on the cityscape dataset,[1] calculating the 14,115 by 19 semantic segmentation matrix. The number 19 represents 19 classes of features with clear physical meanings, namely: road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle. These categories represent the proportion of pixels for each visual element throughout the entire image. The matrices are used to represent the visual features of different scenes in the study area. For example, if a street is long and has multiple panoramic images, the average pooling method is used to characterize the initial features.

As shown in Fig. 5, we divide the study area into several road segments. And we give the road segments in the study area the attributes of traffic status based on the GPS location data of the vehicles. Speed value is calculated by the average speed of the vehicles passing the road segment, and traffic volume is the total number of vehicles passing the road segment within one day. Fig. 6 shows the distribution of the

[1] Please note that you can use any semantic segmentation model according to your needs. Here, we based our approach on the Cityscape classification standard (19 categories), which includes a richer set of urban landscape physical semantics, making it more conducive to interpretable analysis.

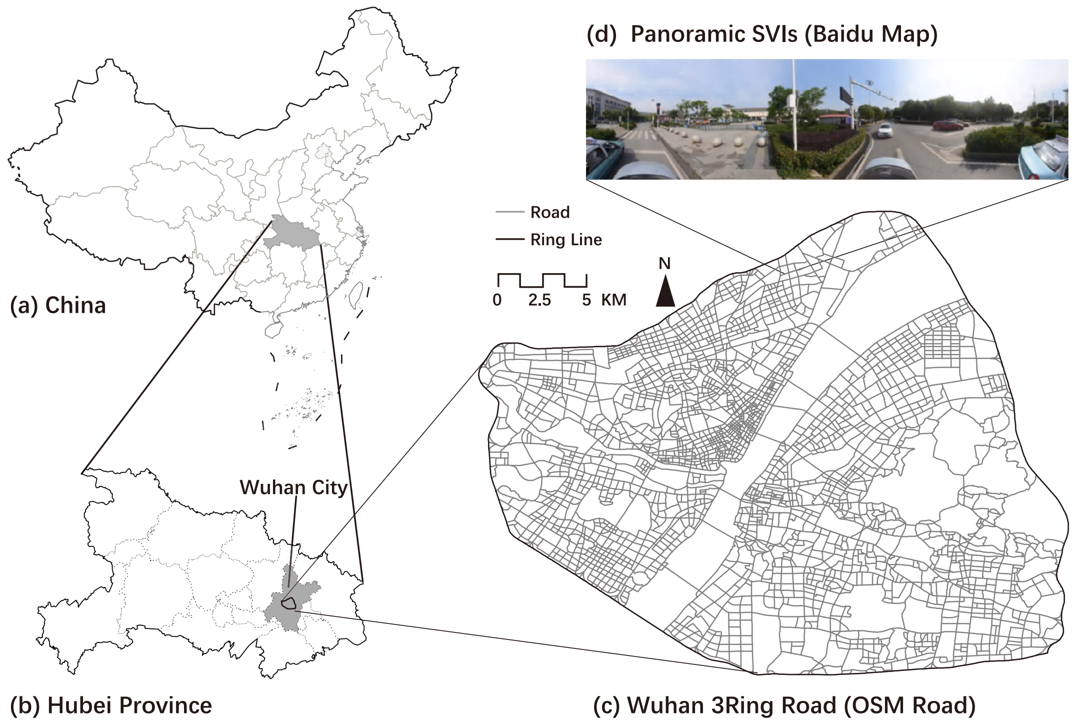

Fig. 5. Schematic diagram of the study area and research data.

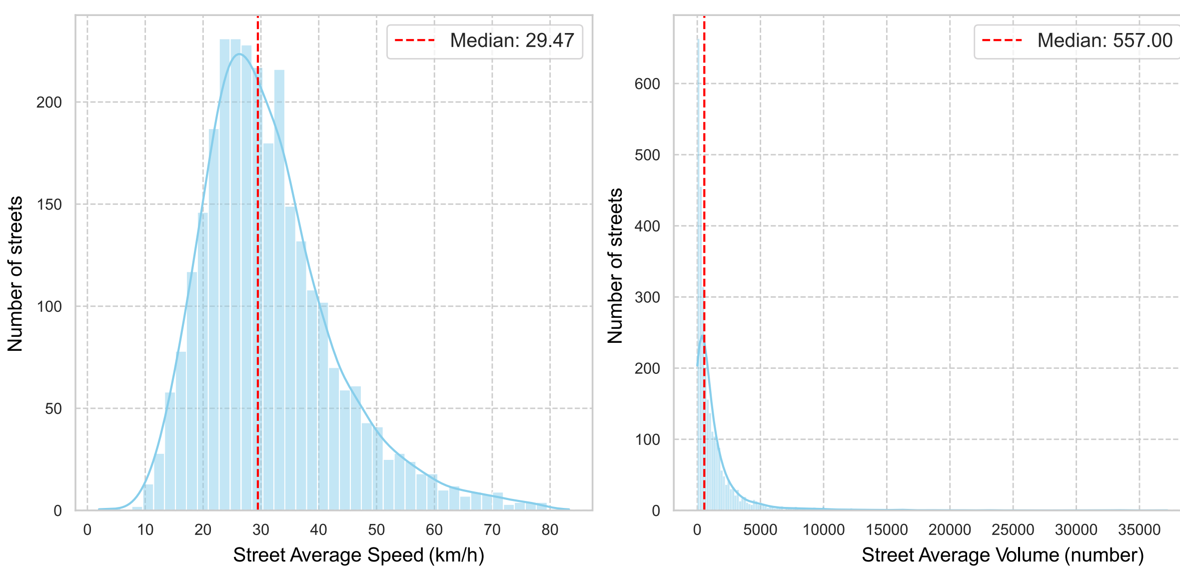

Fig. 6. Comparative histograms with kernel density estimations showcasing the distribution of street-level traffic data. On the left, the average speed distribution with a median of 29.47 km/h, and on the right, the average traffic volume distribution with a median of 557 vehicles. A total of 5075 road samples.

data, the data's distribution is noteworthy, with the speed distribution balanced and conforming closely to a Gaussian distribution. In contrast, the flow volume distribution adheres to a power law distribution, where a minuscule number of nodes are responsible for the majority of traffic flow. The traffic attributes of street are classified into different types using the quantile classification method, ensuring an equal number of samples for each type. Compared to dynamic natural break classification, this method balances the prediction samples. In the meantime, it also increases the difficulty of estimation.

In this paper, short-distance dependence refers to the direct impacts resulting from the proximity or connectivity of locations through roads. Conversely, long-distance spatial interactions refer to the implicit associations between scenes that go beyond physical adjacency. Long-distance spatial interactions are primarily described by human activities,[2] which can be characterized using social media data, GPS mobility data, and sensor data. In this study, we mainly use social media check-in data to quantify long-distance spatial interactions. The fundamental principle behind this data collection is monitoring all social media check-in activities within the research area over a one-year period (2023). Tracking the sequence of check-ins made by individual

[2] Including social activities (e.g., gatherings, meeting friends), work activities (e.g., office work), leisure and recreational activities (e.g., shopping, tourism), transportation and commuting (e.g., daily commutes, long-distance travel), and public events (e.g., sports events, large gatherings).

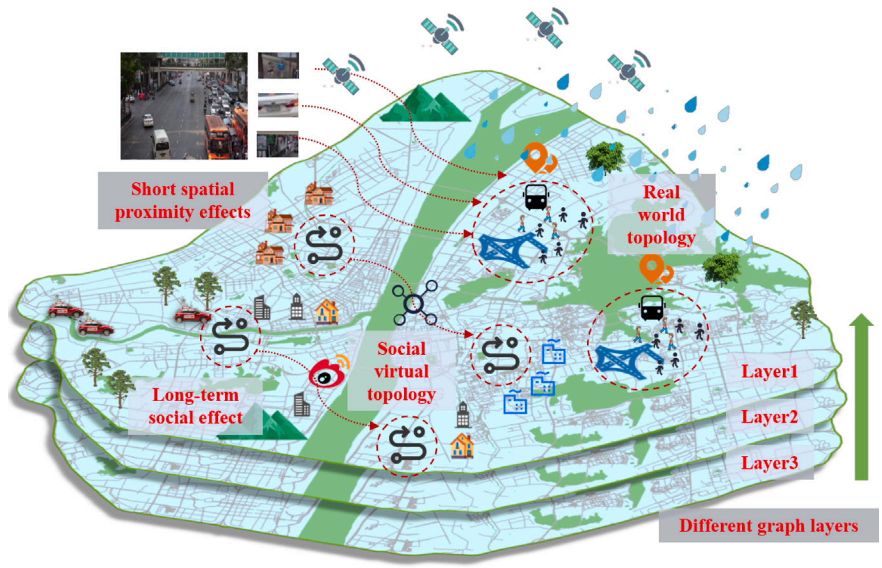

Fig. 7. The illustration of spatial dependence and spatial interaction relationship. Through tracking the sequence of check-ins by individual users to infer latent spatial interactions between streets, reflecting both short spatial dependence effects and long-distance spatial interaction across hybrid graph layers.

Table 2
Comparative and ablation experiment of different algorithms (Three categories classification).

Algorithm	Traffic speed prediction			Traffic volume prediction		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Our method	**0.6034**	**0.6110**	**0.603**	**0.6546**	**0.6600**	**0.6546**
Ablation experiment 1	0.5772	0.5711	0.5772	0.6013	0.6042	0.6013
Ablation experiment 2	0.5620	0.5545	0.5620	0.5495	0.5393	0.5495
GCN	0.5538	0.5484	0.5538	0.5373	0.5320	0.5373
Random forest	0.5481	0.5392	0.5481	0.5229	0.5148	0.5229
Support vector machine	0.5265	0.5300	0.5265	0.5075	0.5002	0.5075
Logistic regression	0.4991	0.4984	0.4991	0.4792	0.4581	0.4792
K-nearest neighbors	0.4683	0.4612	0.4683	0.4585	0.4586	0.4585
Naive Bayes	0.4584	0.4744	0.4584	0.4872	0.4973	0.4872
Decision tree	0.4722	0.4745	0.4722	0.4127	0.4118	0.4127
Gradient Boosting	0.5217	0.5127	0.5217	0.4659	0.4594	0.4659
Neural network	0.5373	0.5341	0.5373	0.4915	0.4710	0.4915

Note: 20% of nodes participate in training, and the remaining 80% nodes make predictions.

users (as shown in Fig. 7), we can infer latent spatial interactions between the locations visited. Concisely, locations that consistently undergo a shared check-in sequence from a user may inherently exhibit spatial interaction.

The processing method we implement initiates a spatial link between the established check-in points and scene segments, clearly depicted as roads in Fig. 5. This is achieved by mapping each check-in to the nearest road in the network, thus creating a connection between human activities and the physical connections. Subsequently, we construct a spatial interaction matrix ($\tilde{W}_2$) based on the spatial behavior sequences of the same user. This approach allows us to reveal patterns that are ignored by traditional traffic analysis methodologies. In the final step, we assemble these spatial interactions to create a comprehensive weighted interaction network constituting 131,652 interaction relationships. Here, each weight mirrors the robustness and frequency of the spatial interactions deduced from social media check-in sequences. The method of creating the spatial dependency matrix is widely recognized and for specific construction methods, refer to such studies as Zhang et al. (2023a).

4.2. Model training process and comparative experimental results

To verify the effectiveness of the method proposed in this paper, we labeled the street according to the quantile method (mentioned in Section 4.1). These streets were categorized into high-, medium-, and low-speed road sections in terms of speed, and into high-, medium-, and low-traffic road sections based on speed traffic flow. Furthermore, we conducted 10 sets of comparative experiments, including the method proposed in this study, traditional GCN network methods (Lee and Rhee, 2022), as well as random forest, support vector machine, logistic regression, K-nearest neighbors, naive Bayes, decision tree, gradient boost, and neural network (Zhang et al., 2018c; Stepanov et al., 2020). In addition, we conducted two ablation experiments: Ablation Experiment 1 excluded the module introduced in Section 3.1, while Ablation Experiment 2 neglected the spatial interaction delineated in Section 3.2. Thus, A total of 12 comparative experiments were performed as shown in Table 2. In order to provide an extensive evaluation of our method's performance, we undertook additional experiments for four-class traffic state estimation alongside the earlier-discussed three-class traffic state prediction. The experimental results under both three-class and four-class labels are tabulated in Tables 2 and 3, with a split of 20% nodes for training and 80% for testing, under the traffic speed and traffic flow, respectively. Our experiment utilized the PyTorch framework. The model architecture featured a consistent hidden dimension of 30 across various layers to ensure uniformity and control over the model's complexity. Key components included a shared Graph Attention Network layer for initial feature processing and multiple independent GAT layers for further feature extraction from

Table 3
Comparative and ablation experiment of different algorithms (Four categories classification).

Algorithm	Traffic speed prediction			Traffic volume prediction		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Our method	**0.5373**	**0.5342**	**0.5373**	**0.5938**	**0.5923**	**0.5938**
Ablation experiment 1	0.5156	0.5080	0.5156	0.5672	0.5743	0.5672
Ablation experiment 2	0.4705	0.4664	0.4705	0.5458	0.5451	0.5458
GCN	0.4531	0.4520	0.4531	0.5373	0.5495	0.5373
Random forest	0.3837	0.3775	0.3837	0.4115	0.4053	0.4115
Support vector machine	0.3924	0.3997	0.3924	0.3966	0.3883	0.3966
Logistic regression	0.3524	0.3511	0.3524	0.3923	0.3837	0.3923
K-nearest neighbors	0.3681	0.3661	0.3681	0.3475	0.3556	0.3475
Naive Bayes	0.3247	0.3352	0.3247	0.3348	0.3296	0.3348
Decision tree	0.3073	0.3068	0.3073	0.3113	0.3102	0.3113
Gradient Boosting	0.3455	0.3441	0.3455	0.3817	0.3769	0.3817
Neural network	0.3750	0.3681	0.3750	0.3881	0.3836	0.3881

Note: 20% of nodes participate in training, and the remaining 80% nodes make predictions.

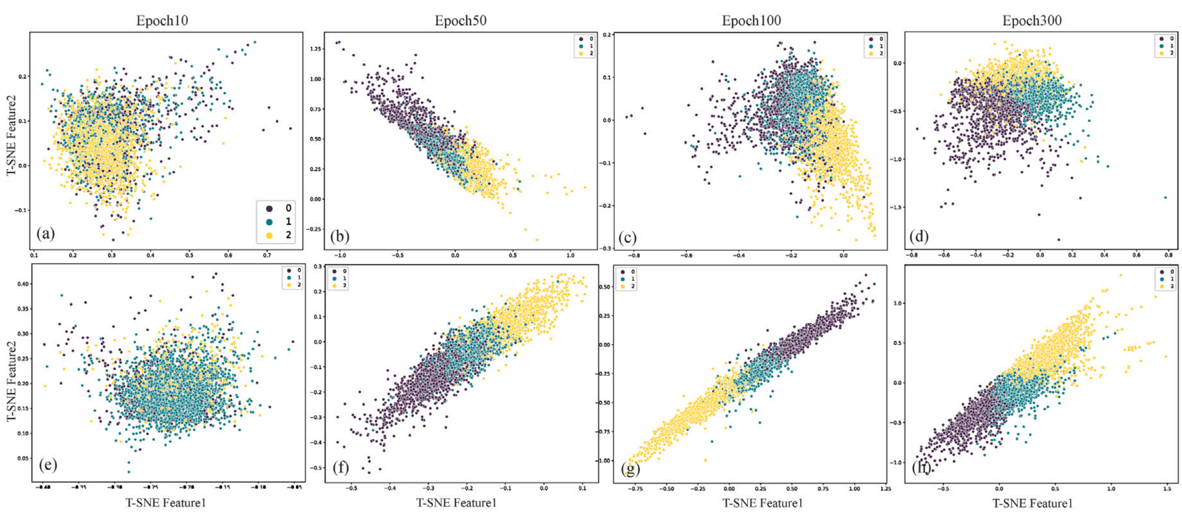

Fig. 8. Evolution of t-SNE clustering over training epochs for traffic volume and speed estimation. Subplots (a) to (d) illustrate the clustering at epochs 10, 50, 100, and 300, respectively, for traffic volume, while subplots (e) to (h) correspond to the same epochs for traffic speed. Each color represents a different class label, indicating distinct traffic conditions. The progression from diffuse to distinct clusters demonstrates the model's increasing accuracy and memory retention in distinguishing between varying traffic patterns, with a more pronounced definition in traffic volume clustering, as indicated by the tighter and more separated groupings, especially in later epochs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

different input axes. Additionally, we incorporated attention mechanisms to enhance local feature processing and used a fully connected layer for classification during the feature fusion stage. In addition, the computations were executed on an RTX4060, significantly accelerating the training process of this complex model. We employed the Adam optimizer to handle the optimization process and set the learning rate to 0.001, dictating the step size during each iteration towards achieving a loss function's minimum. We also integrated dropout and batch normalization techniques to prevent overfitting, enhancing the model's robustness during training. The results were compared using Accuracy and Precision & Recall metrics (Section 3.3).

It is evident that our method consistently outperforms others in both traffic flow and speed estimation across various control groups. This underscores the effectiveness of our designed strategy for the fusion of spatial dependence & spatial interaction information, and highlights how the incorporation of a residual structure significantly enhances the model's capacity for memory retention. Moreover, the model demonstrates superior performance in predicting traffic flow compared to traffic speed. This observation suggests a correlation between the visual features of roads and variations in traffic volumes or road categories. In reinforcing the robustness of our method, this also foreshadows the prospective utility of street-view imagery as a fecund source for traffic flow analysis (Hankey et al., 2021; Tanprasert et al., 2020).

The visualization of t-SNE serves as a tool to examine the features extracted before the classification layer of our model. By reducing these features to a two-dimensional mathematical space, we observe the distribution of different scene features across various training epochs, as shown in Fig. 8. The plots distinctly illustrate that traffic volume is more easily distinguished. Even with fewer epochs, the model is capable of identifying different traffic states. This rapid feature differentiation is indicative of the model's efficiency in learning to distinguish between varying levels of scene congestion. Interestingly, as training progresses, the model's ability to separate these traffic features becomes increasingly pronounced. The initial epochs show a broader blend of features within the space. However, the subsequent epochs reveal a more defined separation, underscoring the model's developing proficiency in recognizing and categorizing traffic volume patterns. In contrast, the model requires more epochs (epoch > 300) to achieve a more significant distinction in the prediction of traffic speed.

In total, the visual progression documented in these plots corroborates our quantitative findings, asserting that the designed residual structures and information fusion structure within our model significantly bolster its memory retention capability, particularly for traffic volume predictions. The ensuing section probes into the spatial attribution analysis across different models, illuminating the spatial distribution of traffic features within diverse urban layouts and street designs.

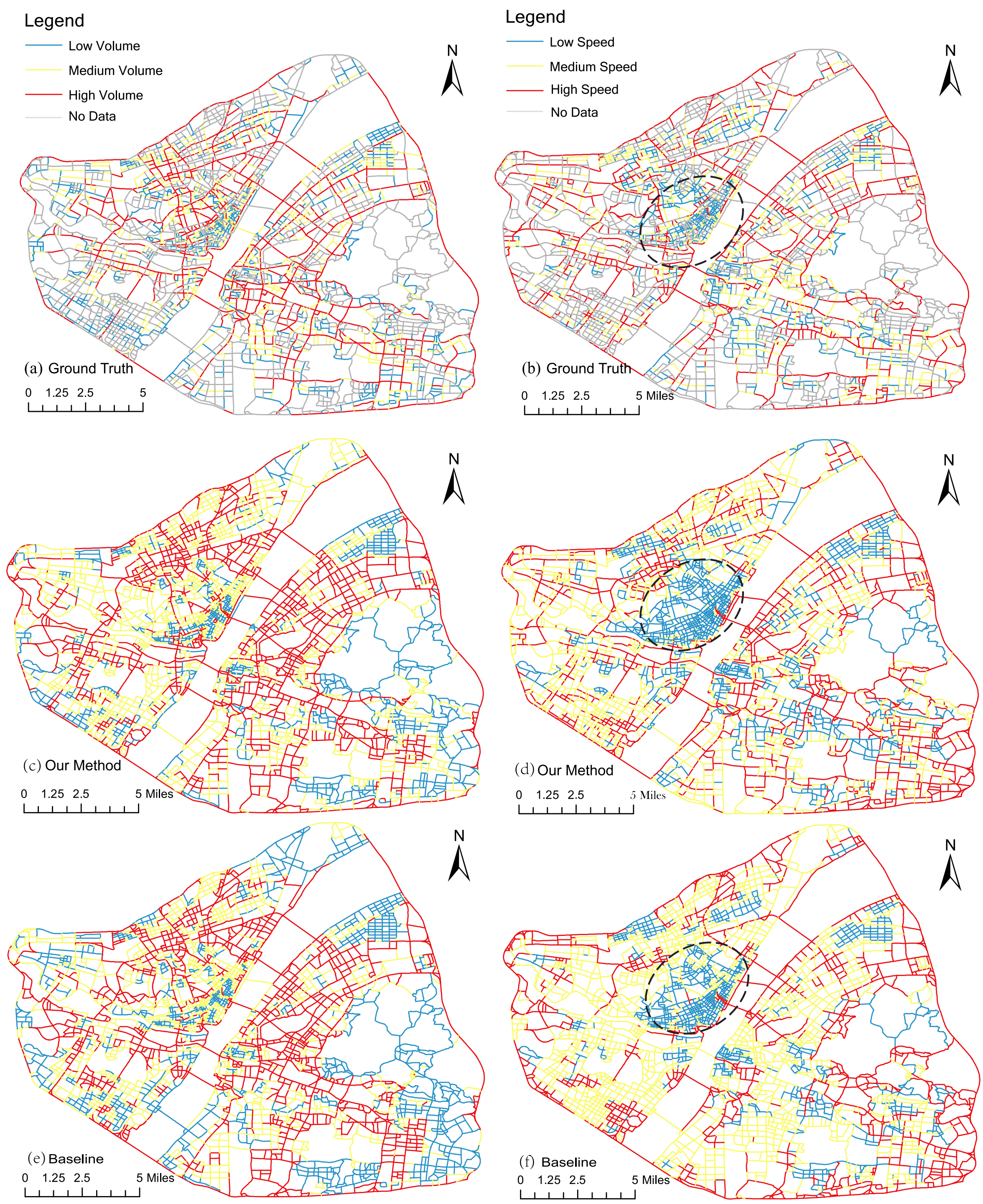

Fig. 9. Comparative visualization of traffic flow volume and speed predictions. Panels (a) and (c) illustrate the ground truth of traffic flow volume and speed. Panels (c) and (d) illustrate the predicted distribution of traffic volume and speed, respectively, after accounting for long-distance spatial interactions. Panels (e) and (f) showcase the spatial distribution of predicted values when the model considers only near-range spatial dependencies. This figure emphasizes the enhanced predictive accuracy in most areas when long-distance spatial interactions are integrated into the model. It particularly highlights the marked decrease in predicted road section flow speeds (panels (d) and (f)) and the significant increase in traffic volume predictions (panels (c) and (e)) in urban core areas, underscoring the potential limitations of traditional traffic forecasting methods in high-density urban zones.

4.3. Spatial attribution analysis of differences in model predictions

To illustrate the spatial performance between various methods, we draw Fig. 9, presenting the predictive outcomes of Ablation Experiment 1 (rely solely on spatial dependency information) or our method (Section 3.2, combining of spatial dependency and spatial interaction information). In Fig. 9, panels a and b demonstrate the traffic flow volume prediction. On the contrary, panels c and d traffic flow speed prediction. In these four comparisons of predictive visualization, we can observe that our model predicts slower road travel speeds compared to traditional methods (panels d and f) in the main urban areas. This disparity could be attributed to our method's higher accuracy in predicting congestion levels within cities due to the consideration of long-distance spatial interactions between roads. Traditional methods

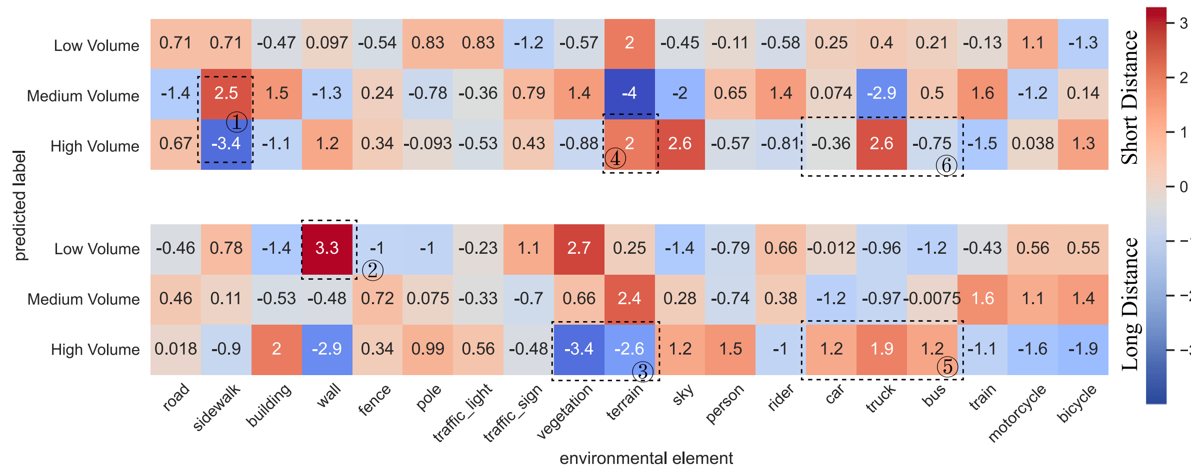

Fig. 10. Heatmap displaying the interpretation influence of various visual elements on traffic volume predictions at the image level. With cell values representing the degree of impact, across categories of low, medium, and high traffic volumes between spatial dependence and spatial interaction efforts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

may underestimate congestion in core urban areas as they do not fully consider inter-regional traffic impacts. This is corroborated by Tables 2 and 3, which show a general incline in prediction improvement across most regions due to the incorporation of long-distance spatial interaction factors. On the flip side, traffic flow predictions from our model and traditional methods do not display significant differences, both achieving compatible prediction levels. Additionally, we present the ground truth of traffic states in Fig. 9, further validating that our estimated results mirror actual traffic conditions.

Overall, our method achieves a precise and detailed representation of urban traffic scenarios, unlike traditional approaches that largely underestimate congestion in dense urban traffic regions. Our method, which underlines the significance of both local range and long-distance spatial interactions, provides a more sound and reliable structure for urban traffic prediction.

4.4. Interpretable validation of model results at different levels

Building on the visual discussion from the previous section, in this section, this section is dedicated to carrying out a lucid and interpretive analysis on intra-street and inter-street scale levels to augment the physical interpretability of our street representation method.

To clearly illustrate how visual elements affect prediction values at the inter-street level, we plotted Fig. 10. The interpretation analysis is conducted separately considering the spatial interaction matrix (Eq. (11)) and the spatial dependency matrix(Eq. (10)), as discussed in Section 3.2. This approach assists in thoroughly understanding how the variations in environmental features under contrasting spatial weights impact the predicted values. This analysis utilizes PyTorch hook technology,[3] which is an advanced feature in the PyTorch library that allows for the extraction and examination of intermediate data at specific layers during the model's forward or backward pass. A "hook" in the context of neural networks is essentially a function that can be attached to a particular layer of a neural network. Once attached, it can monitor and modify the outputs or gradients that pass through that layer. In practice, hooks can be used for a variety of purposes, such as debugging and modifying gradients for custom training procedures, as in our case, for feature importance analysis.

In this heat map (Fig. 10), each cell reflects the change in predicted traffic volume when the corresponding environmental feature is altered. Color variation from blue to red illustrates the scale of influence, with blue indicating a negative influence and red a positive one. Thus, a blue cell correlates an increased value of the environmental feature to a reduction in traffic volume, whereas a red cell suggests an increase in both environmental feature and traffic volume.

We explore the differential impacts of environmental variables on traffic volume predictions, categorizing the effects into two distinct types: short-distance impacts governed by spatial dependence, and long-distance impacts influenced by spatial interactions. The results, depicted in Fig. 10, uncover how variability in these factors affects traffic volume across different spatial scales. Increased sidewalk coverage has significantly decreased short-distance traffic flow, as illustrated in Fig. 10-①. The expansion of pedestrian areas restricts vehicular movement on adjacent streets, directly affecting the flow of traffic within close proximity. In contrast, the impact of structural barriers such as walls becomes more pronounced over long distances. Fig. 10-② indicates a substantial reduction in long-distance traffic volume with an increase in the content of walls. This suggests that such structures disrupt the continuity of routes typically used for longer spatial interactions, thereby influencing the overall connectivity and accessibility across urban scenes. Additionally, an increase in the semantic proportion of buildings correlates with an enhancement in long-distance traffic flow. Areas with high building density frequently serve as hubs of human activity, thereby attracting traffic from extensive regions. The presence of dense building areas facilitates long-distance spatial interactions by providing the necessary services and opportunities that draw people from various distances. The effects of natural features such as vegetation and varied terrain are also noteworthy. These visual elements tend to discourage long-distance travel by creating physical barriers, as evidenced in Fig. 10-③ and ④. Finally, the analysis reveals a correlation between the presence of numerous vehicles (cars, trucks, and buses, Fig. 10-⑤ and ⑥) and increased long-distance traffic flow, suggesting that these scenes are major traffic arteries. The results in Fig. 10 align with our intuitive understanding, indirectly validating the feasibility of inferring the characteristics of urban traffic flow from street-view imagery.

For interpretation at the street level, attention weight scores are crucial to identify and emphasize key features within datasets. The model can 'pay more attention' to critical scenes that could potentially impact surrounding traffic speed or volume by assigning higher weights to them. Fig. 11 visualizes the attention weights of matrix $\tilde{W}_1$ (Eq. (10)) and matrix $\tilde{W}_2$ (Eq. (11)) at the street level, respectively. Here, the reddest connections between streets signal robust mutual influence, while the blue denotes an absence of significant interactions. Fig. 11(a) illustrates the spatial dependency weights between streets, exhibiting a discernible clustering pattern, i.e., red and blue lines frequently appear in groups. Conversely, Fig. 11(b) describes the visualization

[3] https://pytorch.org/.

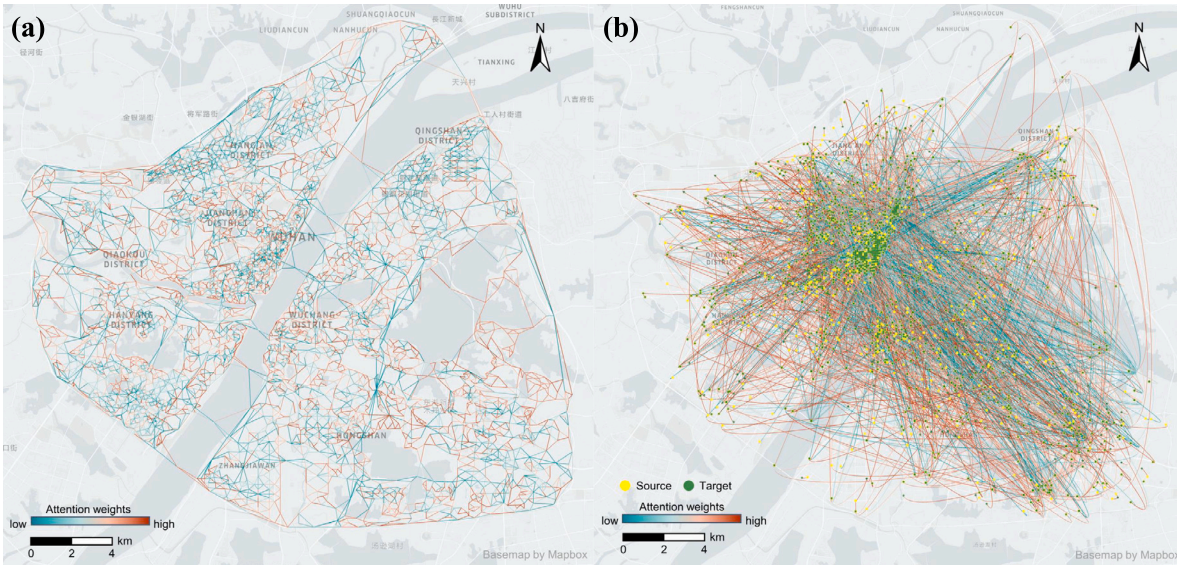

Fig. 11. The visualization of the attention map matrices $\tilde{A}_1$ and $\tilde{A}_2$ at the street level: (a) highlights spatial dependence within the localized street, and (b) illustrates spatial interactions across the street. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the spatial interaction weights between streets, which reveal a unique distribution state such as strong influences over long distances. This particular configuration aids the model in capturing longer-range spatial correlations associated with complex spatial relations.

5. Discussion and conclusion

In conclusion, this study makes two significant contributions to the field of urban visual intelligence and GeoAI. Firstly, we introduce a new method based on graph neural networks that effectively accounts for spatial interactions. This method takes advantage of the power of spatial graph modeling and ensures efficient incorporation of short-distance spatial dependencies deriving from street topology, as well as long-distance spatial interactions extracted from extensive social media check-in data.

Secondly, our local self-attention module could fine-tune the weighting of street features dynamically, enabling adaptation to a variety of downstream tasks. Unlike traditional methods that generate fixed representations (e.g., semantic segmentation-based features that only produce static representations of trees, buildings and vehicles.), our approach adaptively captures the most essential environmental factors according to different scenarios. Moreover, the information fusion through residual connections and a hybrid graph within the graph convolutional layers significantly enhances the model's robustness, ensuring comprehensive feature representation across the entire network architecture. The model has achieved outstanding results in both flow speed and traffic volume estimation tasks, with ablation experiments further confirming the effectiveness of each module.

In addition, our spatial explainable interpretative analysis uncovers a critical insight: traditional graph neural network algorithms tend to underestimate traffic congestion levels in high-density urban centers. Our experimentation has affirmed the efficacy of using street-view imagery and road network topology for urban street representation. This method is also applicable to a broader range of urban downstream tasks, serving as an effective initial visual representation for applications such as population estimation, economic forecasting, and function recognition. However, this study has certain limitations: The Weibo data involved in this study cannot fully represent all social activities. The core of our work is not to comprehensively express all social activities, but rather to consider both the connections between streets in physical space (spatial dependency) and certain social interactions in social space when representing urban streets. Additionally, social activities encompass a wide range of interactions and behaviors that go beyond online check-ins. Our approach may not be able to comprehensively consider all interactions, and we must decide which type of social interaction to incorporate into our framework based on specific contexts and application scenarios.

CRediT authorship contribution statement

Yan Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis. **Yong Li:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Fan Zhang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (42371468) and the High-performance Computing Platform of Peking University (HPC2306190166).

References

Askarizad, R., Jinliao, H., Jafari, S., 2021. The influence of COVID-19 on the societal mobility of urban spaces. Cities 119, 103388.
Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. Landsc. Urban Plan. 215, 104217.
Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS J. Photogramm. Remote Sens. 163, 82–97.
Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A., 2021. A comprehensive survey of scene graphs: Generation and application. IEEE Trans. Pattern Anal. Mach. Intell. 45 (1), 1–26.
Chen, Y., Zhao, P., Lin, Y., Sun, Y., Chen, R., Yu, L., Liu, Y., 2024. Semantic-enhanced graph convolutional neural networks for multi-scale urban functional-feature identification based on human mobility. ISPRS Int. J. Geo-Inf. 13 (1), http://dx.doi.org/10.3390/ijgi13010027, URL: https://www.mdpi.com/2220-9964/13/1/27.
De Sabbata, S., Ballatore, A., Miller, H.J., Sieber, R., Tyukin, I., Yeboah, G., 2023. Geoai in urban analytics. pp. 1–9.

Fan, J., Weng, W., Tian, H., Wu, H., Zhu, F., Wu, J., 2024. RGDAN: A random graph diffusion attention network for traffic prediction. Neural Netw. 106093.
Fan, Z., Zhang, F., Loo, B.P., Ratti, C., 2023. Urban visual intelligence: Uncovering hidden city profiles with street view images. Proc. Natl. Acad. Sci. 120 (27), e2220417120.
Grekousis, G., 2019. Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis. Comput. Environ. Urban Syst. 74, 244–256.
Guan, F., Fang, Z., Zhang, X., Zhong, H., Zhang, J., Huang, H., 2023. Using street-view panoramas to model the decision-making complexity of road intersections based on the passing branches during navigation. Comput. Environ. Urban Syst. 103, 101975.
Han, Z., Xi, Y., Xia, T., Liu, Y., Li, Y., 2023. Devil in the landscapes: Inferring epidemic exposure risks from street view imagery. In: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems. pp. 1–4.
Hankey, S., Zhang, W., Le, H.T., Hystad, P., James, P., 2021. Predicting bicycling and walking traffic using street view imagery and destination data. Transp. Res. D 90, 102651.
Hu, S., Xing, H., Luo, W., Wu, L., Xu, Y., Huang, W., Liu, W., Li, T., 2023. Uncovering the association between traffic crashes and street-level built-environment features using street view images. Int. J. Geogr. Inf. Sci. 37 (11), 2367–2391.
Huang, W., Zhang, D., Mai, G., Guo, X., Cui, L., 2023. Learning urban region representations with POIs and hierarchical graph infomax. ISPRS J. Photogramm. Remote Sens. 196, 134–145.
Kaczmarek, I., Iwaniak, A., Świetlicka, A., 2023. Classification of spatial objects with the use of graph neural networks. ISPRS Int. J. Geo-Inf. 12 (3), 83.
Kang, Y., Zhang, F., Gao, S., Lin, H., Liu, Y., 2020. A review of urban physical environment sensing using street view imagery in public health studies. Ann. GIS 26 (3), 261–275.
Lee, K., Rhee, W., 2022. DDP-GCN: Multi-graph convolutional network for spatiotemporal traffic forecasting. Transp. Res. C 134, 103466.
Li, M., Gao, S., Lu, F., Liu, K., Zhang, H., Tu, W., 2021. Prediction of human activity intensity using the interactions in physical and social spaces through graph convolutional networks. Int. J. Geogr. Inf. Sci. 35 (12), 2489–2516.
Liu, P., Biljecki, F., 2022. A review of spatially-explicit GeoAI applications in urban geography. Int. J. Appl. Earth Obs. Geoinf. 112, 102936.
Liu, C., Song, W., 2024. Mapping property redevelopment via GeoAI: Integrating computer vision and socioenvironmental patterns and processes. Cities 144, 104644.
Ma, H., Xu, Q., Zhang, Y., 2023. High or low? Exploring the restorative effects of visual levels on campus spaces using machine learning and street view imagery. Urban For. Urban Green. 88, 128087.
Meng, H., Huang, X., Mao, X., Xia, Y., Lin, B., Zhou, Y., 2023. The formation and proximity mechanism of population flow networks under multiple traffic in China. Cities 136, 104211.
Pagani, A., Mehrotra, A., Musolesi, M., 2021. Graph input representations for machine learning applications in urban network analysis. Environ. Plan. B 48 (4), 741–758.
Pan, Y., Chen, S., Niu, S., Ma, Y., Tang, K., 2020. Investigating the impacts of built environment on traffic states incorporating spatial heterogeneity. J. Transp. Geogr. 83, 102663.
Pan, J., Lai, J., 2019. Spatial pattern of population mobility among cities in China: Case study of the national day plus mid-autumn festival based on tencent migration data. Cities 94, 55–69.
Salazar-Miranda, A., Zhang, F., Sun, M., Leoni, P., Duarte, F., Ratti, C., 2023. Smart curbs: Measuring street activities in real-time using computer vision. Landsc. Urban Plan. 234, 104715.
Shao, Z., Sumari, N.S., Portnov, A., Ujoh, F., Musakwa, W., Mandela, P.J., 2021. Urban sprawl and its impact on sustainable urban development: a combination of remote sensing and social media data. Geo-Spatial Inf. Sci. 24 (2), 241–255.
Shi, Q., Zhuo, L., Tao, H., Yang, J., 2024. A fusion model of temporal graph attention network and machine learning for inferring commuting flow from human activity intensity dynamics. Int. J. Appl. Earth Obs. Geoinf. 126, 103610.
Stepanov, N., Alekseeva, D., Ometov, A., Lohan, E.S., 2020. Applying machine learning to LTE traffic prediction: Comparison of bagging, random forest, and SVM. In: 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. ICUMT, IEEE, pp. 119–123.
Tanprasert, T., Siripanpornchana, C., Surasvadi, N., Thajchayapong, S., 2020. Recognizing traffic black spots from street view images using environment-aware image processing and neural network. IEEE Access 8, 121469–121478.
Wang, P., Zhang, H., Cheng, S., Zhang, T., Lu, F., Wu, S., 2023a. A lightweight spatiotemporal graph dilated convolutional network for urban sensor state prediction. Sustainable Cities Soc. 105105.
Wang, P., Zhang, Y., Hu, T., Zhang, T., 2023b. Urban traffic flow prediction: A dynamic temporal graph network considering missing values. Int. J. Geogr. Inf. Sci. 37 (4), 885–912.
Wang, Y., Zhu, D., 2024. A hypergraph-based hybrid graph convolutional network for intracity human activity intensity prediction and geographic relationship interpretation. Inf. Fusion 104, 102149.
Wu, A.N., Biljecki, F., 2023. InstantCITY: Synthesising morphologically accurate geospatial data for urban form analysis, transfer, and quality control. ISPRS J. Photogramm. Remote Sens. 195, 90–104.
Xu, Y., Jin, S., Chen, Z., Xie, X., Hu, S., Xie, Z., 2022a. Application of a graph convolutional network with visual and semantic features to classify urban scenes. Int. J. Geogr. Inf. Sci. 36 (10), 2009–2034.
Xu, Y., Zhou, B., Jin, S., Xie, X., Chen, Z., Hu, S., He, N., 2022b. A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method. Comput. Environ. Urban Syst. 95, 101807.
Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., Yin, B., 2021. Deep learning on traffic prediction: Methods, analysis, and future directions. IEEE Trans. Intell. Transp. Syst. 23 (6), 4927–4943.
Yu, W., Wang, G., 2023. Graph based embedding learning of trajectory data for transportation mode recognition by fusing sequence and dependency relations. Int. J. Geogr. Inf. Sci. 1–24.
Zhang, L., Alharbe, N.R., Luo, G., Yao, Z., Li, Y., 2018c. A hybrid forecasting framework based on support vector regression with a modified genetic algorithm and a random forest for traffic flow prediction. Tsinghua Sci. Technol. 23 (4), 479–492.
Zhang, Y., Chen, N., Du, W., Li, Y., Zheng, X., 2021b. Multi-source sensor based urban habitat and resident health sensing: A case study of Wuhan, China. Build. Environ. 198, 107883.
Zhang, B., Cheng, S., Wang, P., Lu, F., 2024a. Inferring freeway traffic volume with spatial interaction enhanced betweenness centrality. Int. J. Appl. Earth Obs. Geoinf. 129, 103818.
Zhang, F., Fan, Z., Kang, Y., Hu, Y., Ratti, C., 2021a. "Perception bias": Deciphering a mismatch between urban crime and perception of safety. Landsc. Urban Plan. 207, 104003.
Zhang, Y., Liu, P., Biljecki, F., 2023a. Knowledge and topology: A two layer spatially dependent graph neural networks to identify urban functions with time-series street view image. ISPRS J. Photogramm. Remote Sens. 198, 153–168.
Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Liu, Y., Batty, M., Ratti, C., 2024b. Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. Ann. Am. Assoc. Geogr. 114 (5), 876–897.
Zhang, M., Tan, S., Pan, Z., Hao, D., Zhang, X., Chen, Z., 2022a. The spatial spillover effect and nonlinear relationship analysis between land resource misallocation and environmental pollution: Evidence from China. J. Environ. Manag. 321, 115873.
Zhang, F., Wu, L., Zhu, D., Liu, Y., 2019. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. ISPRS J. Photogramm. Remote Sens. 153, 48–58.
Zhang, E., Xie, H., Long, Y., 2023. Decoding the association between urban streetscape skeletons and urban activities: experiments in Beijing using Dazhong Dianping data. Trans. Urban Data Sci. Technol. 2 (1), 3–18.
Zhang, Y., Zhang, F., Chen, N., 2022b. Migratable urban street scene sensing method based on vision language pre-trained model. Int. J. Appl. Earth Obs. Geoinf. 113, 102989.
Zhang, F., Zhang, D., Liu, Y., Lin, H., 2018a. Representing place locales using scene elements. Comput. Environ. Urban Syst. 71, 153–164.
Zhang, Y., Zhao, T., Gao, S., Raubal, M., 2023b. Incorporating multimodal context information into traffic speed forecasting through graph deep learning. Int. J. Geogr. Inf. Sci. 37 (9), 1909–1935.
Zhang, Y., Zheng, X., Helbich, M., Chen, N., Chen, Z., 2022c. City2vec: Urban knowledge discovery based on population mobile network. Sustainable Cities Soc. 85, 104000.
Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H., Ratti, C., 2018b. Measuring human perceptions of a large-scale urban region using machine learning. Landsc. Urban Plan. 180, 148–160.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2016. Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 6230–6239, URL: https://api.semanticscholar.org/CorpusID:5299559.
Zheng, Z., Chen, F., Lu, J., Zhou, S., 2023. Exploring the influence of individual daily activity patterns on activity-space segregation. Trans. Urban Data Sci. Technol. 2 (1), 19–38.
Zhou, H., Yuan, J., Yi, D., Jin, S., Zhao, Y., Zhang, Z., Zhao, Z., Zhang, J., 2024. A novel dynamic quantification model for diurnal urban land use intensity. Cities 148, 104861.
Zhu, D., Wang, N., Wu, L., Liu, Y., 2017. Street as a big geo-data assembly and analysis unit in urban studies: A case study using Beijing taxi data. Appl. Geogr. 86, 152–164.