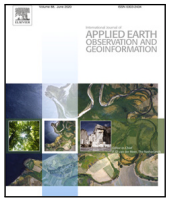
International Journal of Applied Earth Observation and Geoinformation 123 (2023) 103458

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



Inferring socioeconomic environment from built environment characteristics based street view images: An approach of Seq2Seq method☆,☆☆

Yan Zhang [a,b], Fan Zhang [c], Libo Fang [d], Nengcheng Chen [a,b,*]

[a] National engineering research center of geographic information system, China University of Geosciences, Wuhan 430074, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China
[c] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, 999077, Hong Kong, China
[d] Hunan Architectural Design Institute Group Co., Ltd, Changsha, 410208, China

ARTICLE INFO

Keywords:
Social sensing
Seq2Seq
Multi-modal
LSTM
GeoAI

ABSTRACT

The street view image (SVI) and the point of interest (POI) are data sources with different modalities and representing different urban environments, respectively. These two types of data may not fully cover a city. Inferring from one type of data to another will enrich our means of sensing cities and facilitate a deeper understanding of the place. Intuitively, by analyzing visual elements of the urban built environment such as buildings, roads, vehicles, and other features captured in photos of their surroundings, we can infer valuable information about the most likely POI types. Traditional approaches are mainly based on probabilistic statistical models. Such models lack a holistic understanding of the scenario and accurate modeling of the non-linear relationship between the built and socio-economic environments. In response, a Seq2Seq framework is proposed to "translate" and create a bridge between those two data. Experiments in Wuhan demonstrated the high performance of our method (accuracy = 0.770, recall = 0.773, f_1 = 0.771). And this work allows us to better understand the "many-to-many" relationship between the two environments (data), effectively enriching our means of perceiving cities. It also has potential applications in location awareness and visual navigation.

1. Introduction

With the comprehensive coverage of urban space by 5G networks, sensors and digital information, a large amount of multi-source sensing data recording human behavior and urban characteristics is generated especially in cities (Salim et al., 2020; Wang et al., 2022c). These data often exist in unstructured and multi-modal forms, are twin mappings of the real physical environment and help us sense, describe and understand the city.

The modal (O'Halloran, 2011) in this paper refers to the different types of data, such as video, picture, voice, and text, etc. Location-based services provide us with a lot of urban multimodal data with detailed GNSS location information. For example, street view image (SVI) (Zhang et al., 2023), point-of-interest (POI) data (Xue et al., 2020), cell phone signaling data (Sotomayor-Gómez and Samaniego, 2020; Zhang et al., 2022b), geo-tagged flick data (Boeing, 2021), and mobile App data (Huang et al., 2021). They can be organized according to location (Huang et al., 2021), and make us better recognize places (Zhang et al., 2020b).

However, a single data source or a combination of data sources could only reflect a specific problem and is difficult to interpret in relation to each other (Chen et al., 2022). For example, we could easily perceive the "visible" built environment around us (Qin et al., 2020), whether dense trees or crowded buildings. More generally, we could get the geographic entity of the location from the SVI, but that is all. How to infer invisible attributes from visible scenes is still a problem that has not been well solved (Khosla et al., 2014). For example, we do not have access to the types of human activities that exist around the location, which means that most socio-economic attributes are missing (Yin et al., 2021; Chen and Biljecki, 2022). These data would be more meaningful for location-based cognition, decision making, and navigation services. Furthermore, given the possible cognitive bias of unimodal data (Zhang et al., 2021c; Liu et al., 2021), the fusion of

☆ This research was supported by the National Key R&D Program (no. 2018YFB2100500), National Nature Science Foundation of China (nos. 41971351, 41771422) and CAST 2022 Postgraduate Science Popularization Capacity Enhancement Project (KXYJS2022084).
☆☆ The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.
* Correspondence to: 430074, Lumo Road 388, Wuhan, China.
E-mail addresses: sggzhang@whu.edu.cn (Y. Zhang), cnc@whu.edu.cn (N. Chen).
URLs: www.researchgate.net/profile/Yan-Zhang-532 (Y. Zhang), https://grzy.cug.edu.cn/chennengcheng (N. Chen).

https://doi.org/10.1016/j.jag.2023.103458
Received 8 March 2022; Received in revised form 9 August 2023; Accepted 10 August 2023
Available online 30 August 2023

multisource data with multimodal data to understand "place" may have a positive effect on this research (Zhang et al., 2018).

However, in the face of complex non-linear relationship between the built environment and the socio-economic environment, traditional linear models (one-to-one or many-to-one method) lack a holistic understanding of the urban scene. It is difficult to fully map the multidimensional attributes of urban places with a single data model, and it is not possible to maximize the potential for data location alignment (Yang et al., 2020; Zhang et al., 2020a). To address this problem, we introduced the Sequence2Sequence (Seq2Seq) framework, which consists of an encoder and a decoder. This structure has achieved great success (Mohamed et al., 2021) in the fields of machine translation, speech recognition, text summarization, question and answer systems, etc. Compared to traditional statistical learning methods, it supports uncertain length inputs and outputs, avoids complex probabilistic statistics and rule modeling (Feng and Tong, 2020).

In this paper, our focus is on elucidating the translation of these distinct modalities of urban data. POI serves as an abstract representation of locations with elements such as address descriptors, latitude and longitude coordinates, names, and industry categories (Tingting et al., 2020; Yue et al., 2017). It is imbued with human activity and holds significance even beyond geographical context. In contrast, SVI is a fine-grained intuitive expression of the real world (Zhang et al., 2019). It is inexpensive and efficient enough and could be used to obtain key elements and scenes of the city based on location, enriching our approach to environmental sensing (Zhang et al., 2021a).

As mentioned above, POI and SVI both have location information, where POI is considered to reflect socio-economic environment and the other one is considered to be a representative of the built environment. We consider these two types of urban data as different "city language", and build training pairs based on geographical co-occurrence relationships (Chen et al., 2021; Huang et al., 2021). To parse the many-to-many relationships between them, we introduced a cross-modal translation model leveraging the Seq2Seq framework. This innovative approach grants us a fresh perspective for gaining deeper insights into the essence of urban localities.

It could simulate the human visual perception and can recognize the surrounding built environment, as well as the surrounding socio-economic information with the pre-trained sensing enhancement models. Our research makes full use of cross-modal information compared to existing research (Zhang et al., 2022a), which may be helpful in areas such as city-augmented navigation, location-based recommendation, etc.

The main research questions are as follows:

RQ1: What are the relationships between the urban built environment and socio-economic factors, and how are key entities distributed within the city?

RQ2: How can multimodal data integration be harnessed within based on location to facilitate a reciprocal "translation" between environments?

Our study is organized as follows: the second section is the related research of this paper, which introduces the current research status, the third section details our Seq2Seq based method, the fourth section shows our experimental and validation process, and the last section concludes this paper.

2. Literature review

SVI and POI are the most common data sources for urban dynamics sensing and spatial analysis (Du et al., 2020; Deng et al., 2021). We could explore the combination of different POIs to identify urban function (Gao et al., 2017; Liu et al., 2020), and discover the potential function zone co-occurrence relationships (Chen et al., 2020; Tu et al., 2017). It can also be fused with remote sensing image data to obtain a better accuracy of land use classification (Bao et al., 2020; Zhang et al., 2017, 2021b) and to interpret urban morphology (Qian et al., 2020). Similarly, SVIs can also be applied to spatial analysis domains, including 3D visual reconstruction (Cheng et al., 2018), urban greening & urban planning (Cai et al., 2018; Biljecki et al., 2023), soundscape (Zhao et al., 2023), residents' health (Keralis et al., 2020), and transportation (Stiles et al., 2022).

With respect to multi-modal data fusion or inference, we could infer user locations based on semantic information of text (Chen et al., 2021; Zheng et al., 2018); extract location address fused with geospatial feature (Xu et al., 2020); fuse images, texts and geo-maps to implement a holistic textual-visual-geo analysis framework (Shi et al., 2021; Fan et al., 2020).

In addition to discussing multi-modal data, this paper also constructs a sequence based on POI and SVI data representing different environments. Common sequence data are traffic trajectory data (Wang et al., 2022a) and user location data (Stock, 2018). The standard approach is to use long short-term memory (LSTM) to deal with them for classification or prediction (Zhang et al., 2021b). Since some sequences are very long and LSTM will lose a lot of information, a common approach is to introduce the attention mechanism to capture global information (Huang et al., 2019). LSTM is a type of recurrent neural network (RNN), which is computed sequentially (Wang et al., 2022b). In other words, it can only be computed from sequence left to right or from sequence right to left. The computation at time t depends on the computation results at $t-1$, which limits the ability of the model to parallelize. The transformer structure will also be discussed in this study, which is entirely dependent on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution (Zhou et al., 2022).

Urban data is often captured by different sensors in different locations and with different modalities. The multisource, highly dispersed, and heterogeneous properties make them difficult to process, fuse, and analyze. Most studies focus on unimodal data and aggregated analysis of multimodal data, which can lead to data waste and perception bias (Zhang et al., 2021c). In addition, due to the use of simple statistical models (e.g., least squares regression), these methods are inefficient in transforming geographic data into geographic knowledge (Dimakis et al., 2008). For complex urban built-up area scenario, a suitable solution is urgently needed to process (Li et al., 2016; Xu et al., 2021). In this paper, we try to solve "many to many" problem in order to use urban multi-modal data more efficiently, expand the depth of human knowledge about the built environment, and enhance the cognitive ability of the social environment.

3. Materials and methods

As we noted above, the data was obtained from two publicly accessible, globally available datasets: SVI and POI. We sampled more than 20,000 locations in Wuhan city, constructed training pairs, and conducted a comparison experiment using different encoder&decoder (see Fig. 1).

3.1. Location-based training pairs preparation

We obtained SVIs from Tencent Map[1] and POI data from Baidu Map.[2] Both types of electronic map are widely used and equivalent to Google Maps in China. We collected 83,160 images (800*500 pixels) from four angles at $n = 20{,}790$ locations, as shown in Fig. 2.

On the encoder side of the model, we employed the SSD (Single Shot Detection) network for SVI target detection.[3] The SSD model was trained on the Open Images V4 dataset, which contains 14.6M bounding boxes for 600 object classes on 1.74M images, making it the

[1] https://map.qq.com/
[2] https://map.baidu.com/
[3] https://tfhub.dev/google/openimages_v4/ssd/mobilenet_v2/1

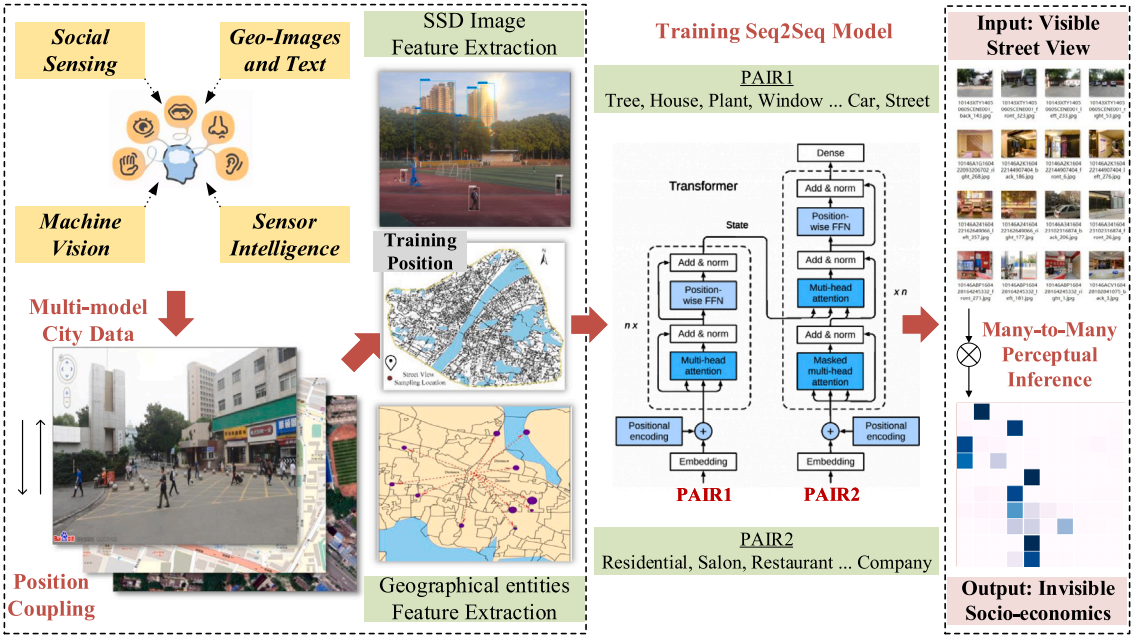


Fig. 1. Technical route of urban sensing enhancement (POI represent the socio-economy and SVI represent the built environment. The extracted urban physical entities are used as the input of the encoder, the socio-economic entities composed of POI lists are used as the output of the decoder, then construct training pairs).

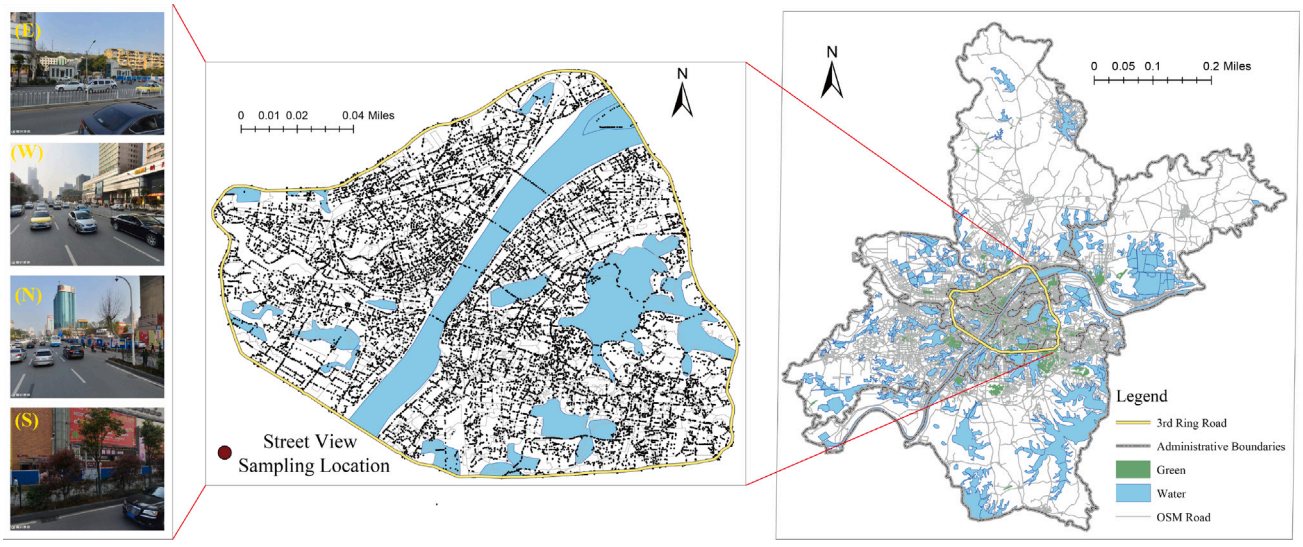


Fig. 2. The SVI sampling location in Wuhan (We choose the main urban area within the main city, including more than 80,000 SVIs in more than 20,000 locations and about 500,000 POI records).

largest existing dataset with object location annotations.[4] Similarly, the decoder side extracts the list of the closest POIs and decodes them by the types.[5]

To expedite model training and attain quicker outcomes, we concentrated on detecting the most evident 100 physical entities within each scene. Moreover, we extracted the 100 closest neighboring POIs as. To represent the location effectively, we extracted a feature sequence

[4] Entity classification for street view target detection, https://storage.googleapis.com/openimages/web/index.html.

[5] The classification reference links of the entities, including the primary, secondary and tertiary classifications of the industry, 22 Basetypes (Caf/Tea, RoadFac, Adr/loc, TourAtr, ShopMal, TrabsFac, Bank/Fina, Sci/Edu, MotServ, CarServ, CarRepa, CarSale, Residen, LivServ, IndoFac, Spr/Rec, ComuServ, Hospital, Gov/Pub, Factory), 220 SUBTYPE (We selected 154 categories that frequently appear in urban for following analysis), and a more detailed 647 CATEGORY by attributes.

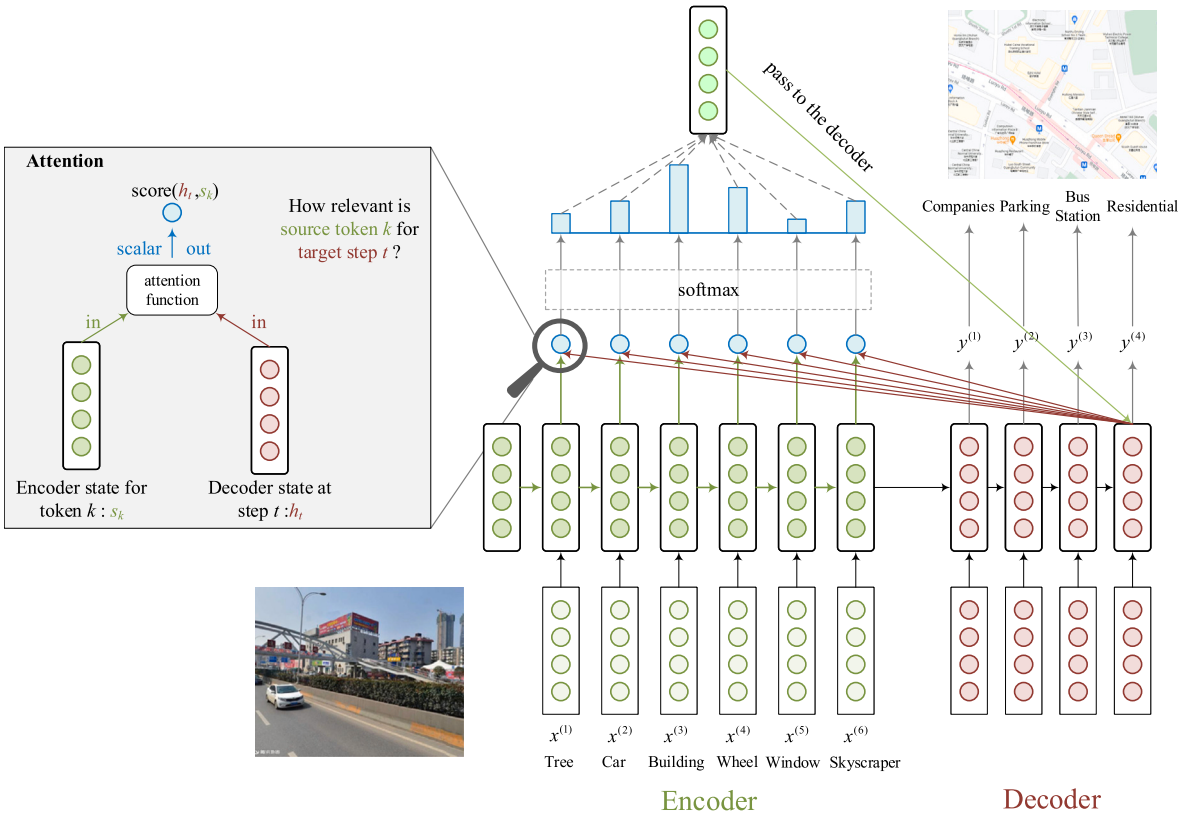


Fig. 3. Seq2Seq with Attention model Schematic (Encoder — reads source sequence and produces its representation; Decoder — uses source representation from the encoder to generate the target sequence.).

based on the highest frequency. This feature sequence, with an input length/output length of 10, captures the most recurrent characteristics. By doing so, we ensured that the model's attention was directed towards the most relevant and significant aspects of each location.

In addition, we construct a total of n sequences Seq_S scene pairs, and tokenize them to build a corpus $Corpus_S$ (in this paper, 139 built environment entities in SVI, e.g., terrain, car, building, etc.). Similarly, we get n sequences Seq_P on the decoder side, and then build the corpus $Corpus_P$. The smallest unit (token) in $Corpus_P$ is POI type (in this paper, POIs are categorized by industry, encompassing 22 BaseType classifications and 154 SUBTYPE classifications,[6] e.g., residential areas, libraries, retail stores, etc.).

During the process of constructing Seq2Seq training pairs, we gathered a total of n (n = 20,790) SVI sampling locations and sought out neighboring POIs. With a substantial m (m = 500,634) POI records available in the city, employing the traditional retrieval method would require an overwhelming $m \times n$ distance queries. This could put a significant strain on computer computing power.

So we construct KD trees to improve the efficiency of retrieval. The computational complexity can be reduced to $O\left(\log^2 m\right) + n \times O\left(m^{1-1/d} + topK\right)$, where $O\left(\log^2 m\right)$ is the computational complexity when building the KD tree. $n \times O\left(m^{1-1/d} + topK\right)$ is the computational complexity of the query, $topK$ is the number of nodes nearest to the query.

Using KD trees, the burden of the retrieval process is significantly alleviated, leading to more efficient and faster results. With d taking the value of 2 (representing longitude and latitude), this approach enables us to effectively find and process the $topK$ nearest nodes to the query locations without compromising the accuracy of the retrieval.

[6] https://lbsyun.baidu.com/index.php?title=open/poitags

3.2. Seq2Seq model

Seq2Seq is a versatile technology framework that has revolutionized traditional fixed-size input and output approaches. It harnesses the power of neural network models for various tasks; one of the most prominent applications is machine translation, where it enables seamless conversion of text from one natural language to another. In this paper, we introduce the encoder processes the input entities, creating a context vector that captures the place's meaningful representation. The decoder then generates the output sequence one POI type at a time, observing the context vector and previously generated POIs. To achieve this, we utilize different encoder/decoder architectures within the Seq2Seq framework, including RNN, RNN-Attention, LSTM, LSTM-Attention, and Transformer. By conducting an accuracy comparison among these variants (Kim et al., 2021), we can identify the most suitable one for our task.

Formally, our input sequence Seq_S consists of elements $(x_1, x_2, \ldots, x_m)$, while the output sequence Seq_P contains elements $(y_1, y_2, \ldots, y_n)$. As depicted in Fig. 3, the Seq2Seq model is essentially a combination of two RNN/GRU/LSTMs: one for data encoding and the other for data decoding. At each timestep t, the encoder's output is fed into the decoder, and following the softmax classifier, the model predicts the probability distribution of all tokens.

$$p^{(t)} = p(* \mid y_1, \ldots, y_{t-1}, x_1, \ldots, x_m) \tag{1}$$

The token with the highest probability is used as the output of the decoder (y_i). The loss function of the model is:

$$Loss(p^*, p) = -\log(p_{y_t}) = -\log(p(y_t \mid y_{<t}, x)) \tag{2}$$

The LSTM is a specialized type of RNN that incorporates three crucial gates: the forgetting gate, the input gate, and the output gate.

This unique architecture has gained widespread recognition for its effectiveness. The LSTM's key strength lies in its ability to regulate the flow of information, enabling the network to selectively retain or discard information as needed. At the core of an LSTM cell, the input is composed of three essential components: the hidden state h_{t-1} and the cell state C_{t-1} from the previous time step, as well as the input x_t at the current time step. These components are combined using specific mathematical formulations, as shown in Eq. (3):

$$\begin{aligned} i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\ f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\ g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\ o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{3}$$

where employs the sigmoid activation function σ and the Hadamard product $\odot$. It comprises weight matrices, such as W_{ii}, representing the weights from the hidden state to the hidden state, and W_{hi}, representing the weight matrix between the input and the hidden state. The primary parameter learned by the LSTM model is W_{hi}, which plays a crucial role in the network's ability to learn and capture sequential patterns.

In order to capture long-term dependencies more effectively, attention mechanism was also introduced. It enables the model to focus on more relevant parts of the input sequence, rather than compressing the entire source into a single vector. It provides representations for all source tokens, and takes all RNN states into consideration instead of just the last one. In the case of the self-attention-based Transformer, at each decoder step, the model determines which source parts are more important for generating the output. This allows the encoder to provide representations for all tokens at different time steps. The attention mechanism then considers all encoder states at the t moments, automatically filtering and assigning weights to important information, which yields attention scores for all encoder states. These attention scores are computed as follows:

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q refers to the query, which corresponds to the content h_t of the decoder at the time step t; K refers to the key, which corresponds to the embedded representation s_k of the output of the source sequence after encoding; V represents the original representation of the source sequence; d denotes the vector dimensions K and V. Softmax operation ensures that the attention scores sum up to 1, allowing the model to attend to relevant information effectively while generating the output sequence.

The Transformer model, while also employing the encoder–decoder structure, diverges from the RNN-based architecture depicted in Fig. 3. Instead, it leverages attention mechanisms to effectively capture and retain semantic information over longer distances. The transformer's encoder and decoder components consist of multiple layers, with each layer incorporating two sublayers: the Multi-Head Attention Layer and the Feed Forward Layer. Furthermore, both sublayers utilize Residual Connections and Normalization to facilitate smoother information flow and alleviate the problem of vanishing gradients. The final output of the sublayer can be expressed as:

$$\text{Sublayer}_{i+1} = \text{LayerNorm}(x + (\text{SubLayer}(x))) \tag{5}$$

where $SubLayer(x)$ denotes the output of each sublayer. Indeed, the Transformer model introduces a significant improvement in computational efficiency compared to traditional RNNs because of its parallel computation capabilities. RNNs process input sequences sequentially, but the Transformer can simultaneously attend to all positions within a sequence, resulting in faster training and inference times. However, one challenge is the lack of inherent positional information for tokens within a sequence. Since attention mechanisms do not consider token positions, it becomes essential to incorporate positional information into the model. To address this, positional encoding is introduced as part of the training parameters. The positional encoding formula is as follows:

$$\begin{cases} p_{k,2i} = \sin\left(k/10000^{2i/d}\right) \\ p_{k,2i+1} = \cos\left(k/10000^{2i/d}\right) \end{cases} \tag{6}$$

Where $p_{k,2i}, p_{k,2i+1}$ are the $2i$, $2i+1$ components of the encoding vector at position k, respectively, and d is the dimension of the position vector. The input sequence is summed by the word embedding and positional embedding and input to the encoder for semantic encoding representation. The Multi-Head Attention layer in the Encoder part maps the input content to different feature spaces to extract different semantic information. All is summed and normalized by Add&Norm, then passed to the full linked feedforward neural network, which outputs the final semantic vector matrix. The output of the lth structure in the encoder is calculated as follows:

$$h_s^{(l)} = Norm\left(MultiHead\left(Q^{(l)}, K^{(l)}, V^{(l)}\right) + Q^{(l)}\right) \tag{7}$$

$$h_{el}^{(l)} = Norm\left(PFF\left(h_S^{(l)}\right) + h_S^{(l)}\right) \tag{8}$$

$$PFF = \left(\text{Re}lu\left(h_s^{(l)}w_s^{(l)}\right)h_s^{(l2)} + h_s^{(l)}\right) \tag{9}$$

where $h_S^{(l)}$ is the result of Multi-Head Attention and residual concatenation, $Norm(\cdot)$ denotes the normalization operation, the output of Encoder l is $h_{el}^{(l)}$, the input of the first structure of encoder is $Q^{(l)} = K^{(l)} = V^{(l)} = E$, and the input of the rest of structures is $Q^{(l)} = K^{(l)} = V^{(l)} = h_{el}^{(l-1)}$, $w_s^{(l)}$ are the training parameters of the model, and PFF denotes the feedforward network layer. The decoder part of the model also consists of a 6-layer structure, and the structure of each layer is similar to the encoder part, which is not repeated here (Vaswani et al., 2017).

3.3. Result accuracy evaluation

For model accuracy evaluation, manual inspection is time-consuming and expensive. So we use the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), which is commonly used in machine translation. The rouge-score is calculated with the following formula:

$$\begin{aligned} &\text{ROUGE (candidate, reference)} \\ &= \frac{\sum_{r_i \in \text{ reference}} \sum_{1-\text{gram} \in r_i} \in \text{reference Count(1-gram, candidate)}}{\sum_{r_i \in \text{ reference}} \text{num1-gram } (r_i)} \end{aligned} \tag{10}$$

where r_i represents the sentence in the reference document (true value), $Count$ denotes the number of times a particular token appears in the candidate document (predicted value); $num1-gram$ denotes the number of times the token (1-gram) appears in the reference document sentence r_i. Depending on the needs of the experiment, the rouge metric also calculates accuracy in the form of token pairs (two tokens, 2-grams); in this paper we use 1-gram rouge, i.e., one token as the smallest unit, to perform the calculation through f_1, $Recall$, and $Precision$.

4. Results

4.1. Spatial and frequent correlation analysis

To investigate the possible spatial correlation between POI and SVI, to assess the feasibility of enhanced sensing, we performed a frequent item extraction and spatial cluster analysis on various environmental entities. The word cloud map presented in Fig. 4 reveals interesting patterns. The categories of socio-economic entities appear to be more

(a) Key entities of the built environment (b) Key entities of the social environment

Fig. 4. Word cloud diagram of key entities in the urban built environment and socio-economic environment.

Table 1
Distribution of urban physical entities and socio-economic entities in different urban topics. (By calculating the perplexity, we identified 11 topics. The five entities with the highest probability in each topic were extracted, and their frequency of occurrence was also calculated).

Topic	topic0	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10
Token	Porch	dormitory	Accommodation services	company	Cold drink shop	Automobile maintenance	famous scenery	Tree	Street	Company enterprise	Government and social organizations
Prob	0.222	0.123	0.194	0.135	0.083	0.127	0.197	0.066	0.073	0.156	0.275
Frequency	1245	2743	1395	6738	801	2409	1306	20692	7873	6357	1153
Token	Flowerpot	Colleges and Universities	Hotel Guest House	Company enterprise	Internet Bar	Automotive service	comfort station	Living service place	Entr and exit of parking lot	Sci,edu and cultural sites	Business housing
Prob	0.077	0.112	0.083	0.113	0.065	0.109	0.102	0.062	0.069	0.121	0.212
Frequency	397	2274	2081	6357	798	2013	1323	18258	1006	2566	882
Token	Houseplant	Building door	Hotel	Front gate of street yard	Clothing	Automobile maintenance	Public Parking	Window	Intersection name	company	Entr and exit of parking lot
Prob	0.061	0.105	0.068	0.084	0.063	0.101	0.085	0.061	0.062	0.121	0.167
Frequency	306	2286	1933	1757	1893	1768	7339	18066	1466	6738	761
Token	Palm	Scientific institution	company	Plant	Catering related	Auto parts sales	scenic spot	Building	Public Parking lot	Financial institutions	Swimming
Prob	0.058	0.078	0.066	0.054	0.058	0.079	0.069	0.059	0.058	0.091	0.101
Frequency	921	1728	6738	13523	4537	1332	455	17512	7339	1299	259
Token	Door	school	driver's school	Logistics Express	Footwear	Automobile maintenance	Parking lot	Car	Bus station	Business office	Government agencies
Prob	0.042	0.064	0.060	0.052	0.053	0.076	0.054	0.059	0.055	0.047	0.075
Frequency	1486	1408	1812	4994	5004	2347	820	17223	4011	709	372

Note: Socio-Economic Environment Entities ● ; Built Environment Physical Entities ●.

evenly distributed, as depicted in Fig. 4b. On the other hand, the categories of physical entities show more concentration, with numerous scenes featuring common visual elements often found in cities, such as buildings, cars, windows, pedestrians, etc., as illustrated in Fig. 4a. This suggests that physical entities tend to exhibit certain visual characteristics that are more prevalent across various locations in the city. In contrast, socio-economic entities seem to display a more diverse distribution, possibly reflecting a broader range of factors and attributes associated with different urban areas.

To perceive the spatial distribution state of urban entities, we plotted the street-visible density of skyscrapers, cars and pedestrians. As shown in Fig. 5. Skyscrapers are detectable in 7281 sampling locations and pedestrians in 9361 sampling locations in the study area. We interrupt the OSM (Open Street Map) road network of Wuhan into 55,032 small segments, each of which is assigned a line density value.

The spatial distributions of skyscrapers, cars, and pedestrians exhibit noticeable differences. To investigate the existence of spatial dependencies between these entities, we employ the Latent Dirichlet Allocation (LDA) Word-Topic-Document toolkit (Sievert and Shirley, 2014) developed by Stanford University. With this toolkit, we conduct frequent co-occurrence term mining of physical and socio-economic entities.

As mentioned earlier, we have gathered n entity pairs, each comprising two sequences. We treat each pair as a "sentence" and analyze the dependencies, or topics, between the entities. By identifying frequent terms that co-occur between different environment entities, we can verify if one type of data can be inferred from the other. Conversely, the absence of frequent co-occurring terms would indicate that the two types of data are independent and cannot be reliably inferred from each other.

Through this analysis, we aim to uncover any potential relationships and dependencies between different types of urban data, shedding light on the interactions and interplay of physical and socio-economic entities within the urban landscape.

We determined the optimal number of clusters (11 topics) using perplexity calculation, as referred in the literature (Gao et al., 2017). The training results are summarized in Table 1. For each topic, we listed the five tokens with the highest probability of occurrence, along with their corresponding probability and frequency values.

In the case of considering only the first five feature tokens, Topic 0 contains only physical entities, and Topics 1, 2, 5, 6, and 9 contain only socioeconomic entities. In addition, there are some other topics including both types of entities, which is our concern.

In addition, we identified 11 different location types. For instance, Topic 0 is indicative of a home residence, with the feature "Porch" having the highest occurrence probability of 0.222, which is the most prominent characteristic of this topic. Topic 1 seems to represent a scientific institution, possibly a university. Meanwhile, Topic 2 is associated with urban hotels and accommodations. Topic 3 conveys a gathering place for companies. Topic 4 may be related to catering services. Topics 5 and 6 are respectively related to car repair and sightseeing tourism themes. The tokens under Topic 7 have approximately equal weights of around 0.06 and lack distinctive characteristics, mainly encompassing physical entities commonly found in urban areas such as roads. Topics 8 and 10 appear to be closely linked to parking lots and streets, respectively. Topic 9 is more focused on companies in the

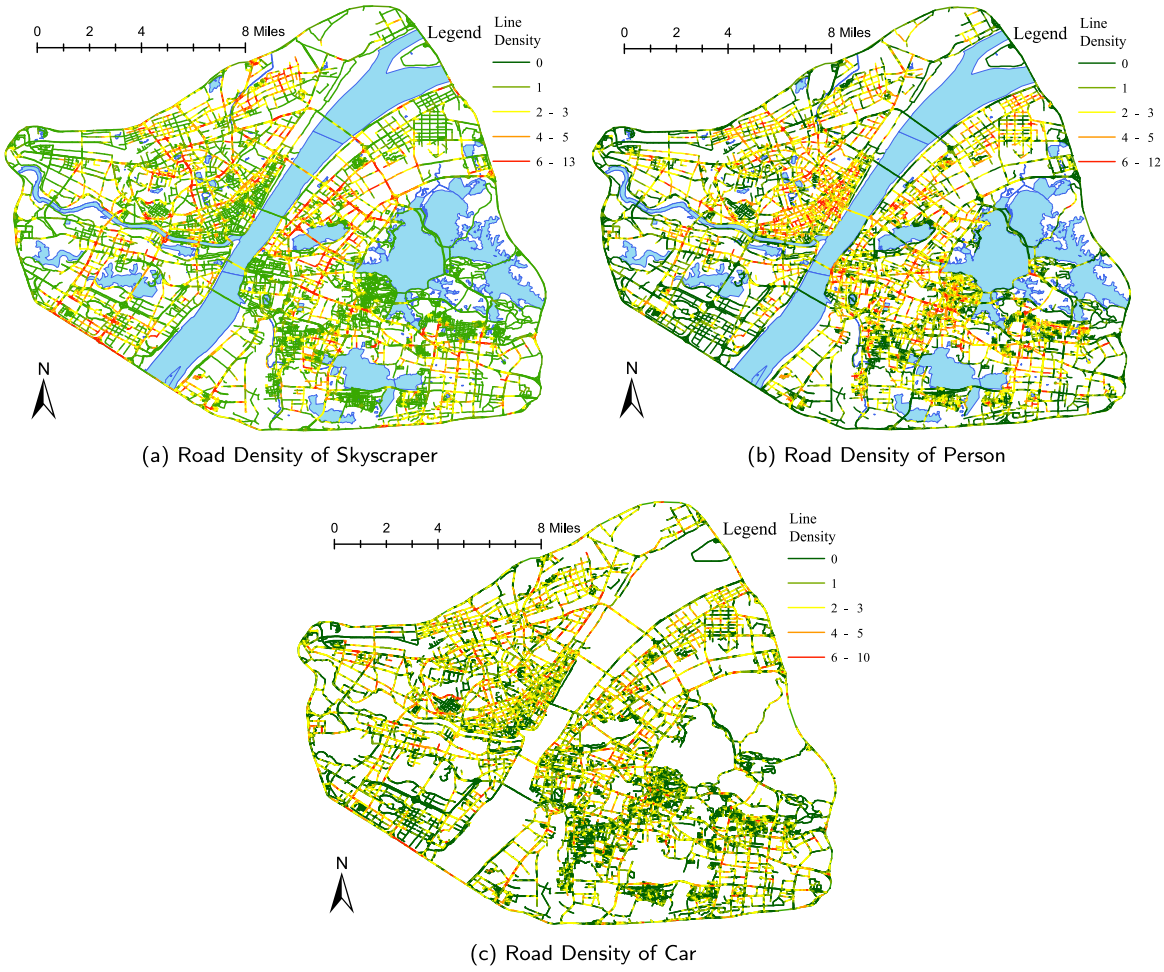


Fig. 5. Sensing the road visible density of skyscraper, person and car through Street View images (We use the OpenStreetMap road network as the base map).

tertiary sector (e.g., finance) than Topic 3. Finally, Topic 10 represents government and administrative agencies.

In this section, we used mostly unsupervised methods. In the next section, we will use the pairs for data translation training and mine many-to-many relationships.

4.2. Training and comparison

To enhance the multi-modal sensing capability, we designed a data translation model mimicking machine translation, and connects the semantic information of POIs to the image features of SVIs.

The SSD is used as a feature extractor of the streetscape images to identify the physical entities and get the sequence Seq_S. On the other hand, POI constructs a KD tree to quickly query the socio-economic entities around the sampled location, and constructs the sequence Seq_P based on the distance.

The experiments were conducted on an NVIDIA Tesla K80 Graphics Processing Unit (GPU) using the Pytorch framework. We set the parameters similar to the settings in the paper (Vaswani et al., 2017), with 8 Self-Attention heads (similar to the convolution kernels in CNN networks, to capture different features of the data).

For the comparison test, we selected common RNN, LSTM, RNN with attention, and LSTM with attention as encoder/decoder components to construct the Seq2Seq models. Two sets of experimental data were created based on the BaseType (base industry type) and SubType (further subdivided industry type) of POIs.

To evaluate the model efficiency, we used three indicators: Rouge-1 f_1 score, Rouge-1 recall, and Rouge-1 precision. These metrics provide a comprehensive assessment of the model's performance in generating accurate and the results are presented in Table 2.

Table 2
Comparison of model accuracy results in SubType and BaseType training data. (We used two scales of data, and five different Encoders & Decoders to compare).

Encoder&Decoder	Rouge-1 *Precision*	Rouge-1 *Recall*	Rouge-1 f_1	Data Type
LSTM-ATTE	0.476	0.470	0.473	●
RNN-ATTE	0.408	0.408	0.408	●
RNN	0.466	0.396	0.417	●
LSTM	0.509	0.461	0.484	●
Transformer	0.499	0.500	**0.499**	●
LSTM-ATTE	0.718	0.662	0.688	●
RNN-ATTE	0.705	0.706	0.705	●
LSTM	0.760	0.690	0.723	●
RNN	0.720	0.704	0.709	●
Transformer	0.770	0.773	**0.771**	●

Note: ● represents SubType; ● represents BaseType.

In summary, the Transformer-based model achieved the best results, outperforming the LSTM-based model by 5% in terms of accuracy (f_1 score). Surprisingly, the addition of the Attention layer did not have a significant improvement (LSTM-ATTE), which could be attributed to the simplicity and shorter length of the "urban language" data. The Transformer could capture complex patterns and dependencies within sequences, making it highly effective in urban environments.

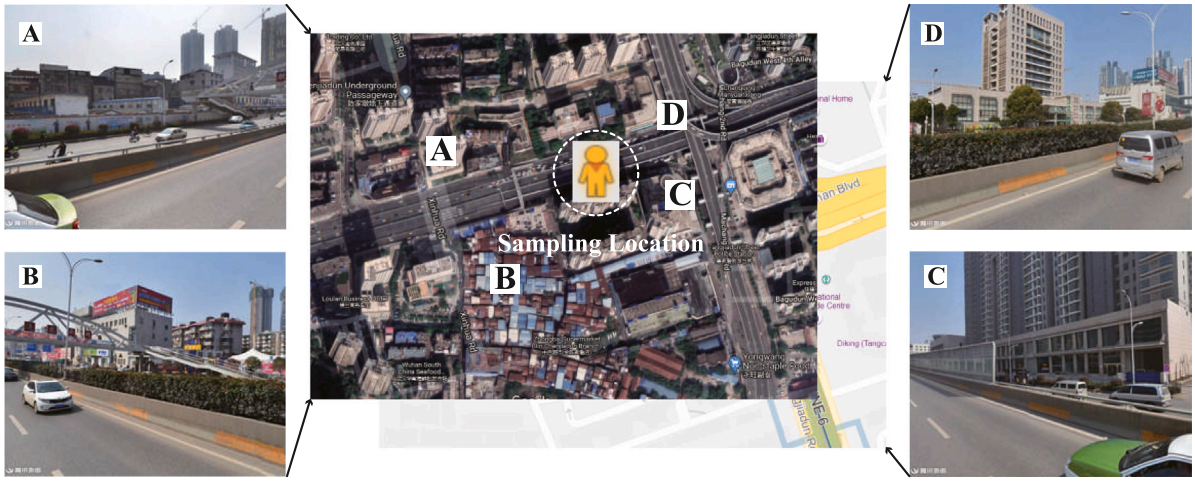


Fig. 6. Street view images of the selected location from four different viewing angles (30.620 N,114.263 E).

Fig. 7. Results of attention heatmap calculation for this scene (The attention score between entities changes in different scenarios).

4.3. Model application

To further illustrate the correlation between those two environments, we generated an attention heat map(Fig. 7). For this demonstration, we chose a specific location (30.620 N, 114.263 E) and used the model to automatically resolve the sequence of physical entities (street view entities) associated with that location, such as Window, Tree, Building, Car, Skyscraper, Wheel, Street Light, House, Vehicle, and Land Vehicle. These entities are shown in Fig. 6.

Following the identification of the physical entities, the model then inferred the most likely surrounding socioeconomic environment, producing an output sequence (predicted geographic entities). This analysis and visualization provide valuable insights into the relationship between the built environment and the socio-economic environment at the selected location. By leveraging the attention mechanisms, the model can effectively capture and highlight the relevant features that contribute to the multi-modal sensing and translation capabilities, enriching our understanding of urban landscapes and their interconnectedness.

Actual socio-economic environment sequence: company, living service, residential area, Chinese restaurant, clinic, agriculture, forestry and fishery base, traffic vehicle control, street gate, government office, food service.

Predicted socio-economic environment sequence: company, street gate, residential area, living service, Chinese restaurant, government office, parking lot, agricultural, forestry and fishery base, bus station, <EOS> (identifier representing the end of the sequence).

It can be seen that the model captured some connections between the two environments. After manual comparison and verification, there are actually parking lots and bus stops in the region as well. It is limited by the feature probability and the length of the sequence. The model does not learn the feature sufficiently and does not output all the entities.

As a comparison, we also provided the co-occurrence frequency of physical entities and socio-economic entities in this scene, as shown in Table 3, while Fig. 7 reveals the attention scores learned by the model for each entity. Skyscraper has a high attention score with restaurant and bus station, and the model believes that there is a strong correlation between them (Zhu et al., 2017b,a). In contrast, this strong correlation is not evident in the Table 3. This proves that the model learns the features in the scenario and performs a personalized computation. Similarly, higher attention scores are observed between Cars and street gates, government agencies, and parking lots. The model also assigns higher attention scores between windows and residential and living service areas, which aligns with our common sense understanding.

The section's research demonstrates that changes in the built environment lead to diverse socio-economic scenarios, but reality is much more intricate. There exists a complex interplay between this two urban environments, characterized by intricate nonlinear relationships that transcend a simple cause-and-effect paradigm. Understanding this relationship is of paramount importance, as it enriches our perception of cities and contributes to a more comprehensive analysis of urban socio-economic patterns.

5. Discussion and conclusion

In this paper, we establish many-to-many relationships between different data sources at the same location and develop a Seq2Seq based sensing enhancement model. It could encode data from two modalities, enabling perception of both environments separately. Users can capture photos of their surroundings to obtain comprehensive location information, independent of electronic maps and navigations. This approach can also be utilized in urban analysis, parsing geo-tagged Twitter or Flickr images to gather socio-economic attributes. The calculated information can be applied to various fields, including urban planning and social computing. Moreover, the model allows the retrieval of similar visual scenes based on surrounding POIs.

In other words, this approach aligns well with advances in foundational models (Mai et al., 2023; Bommasani et al., 2021). By utilizing

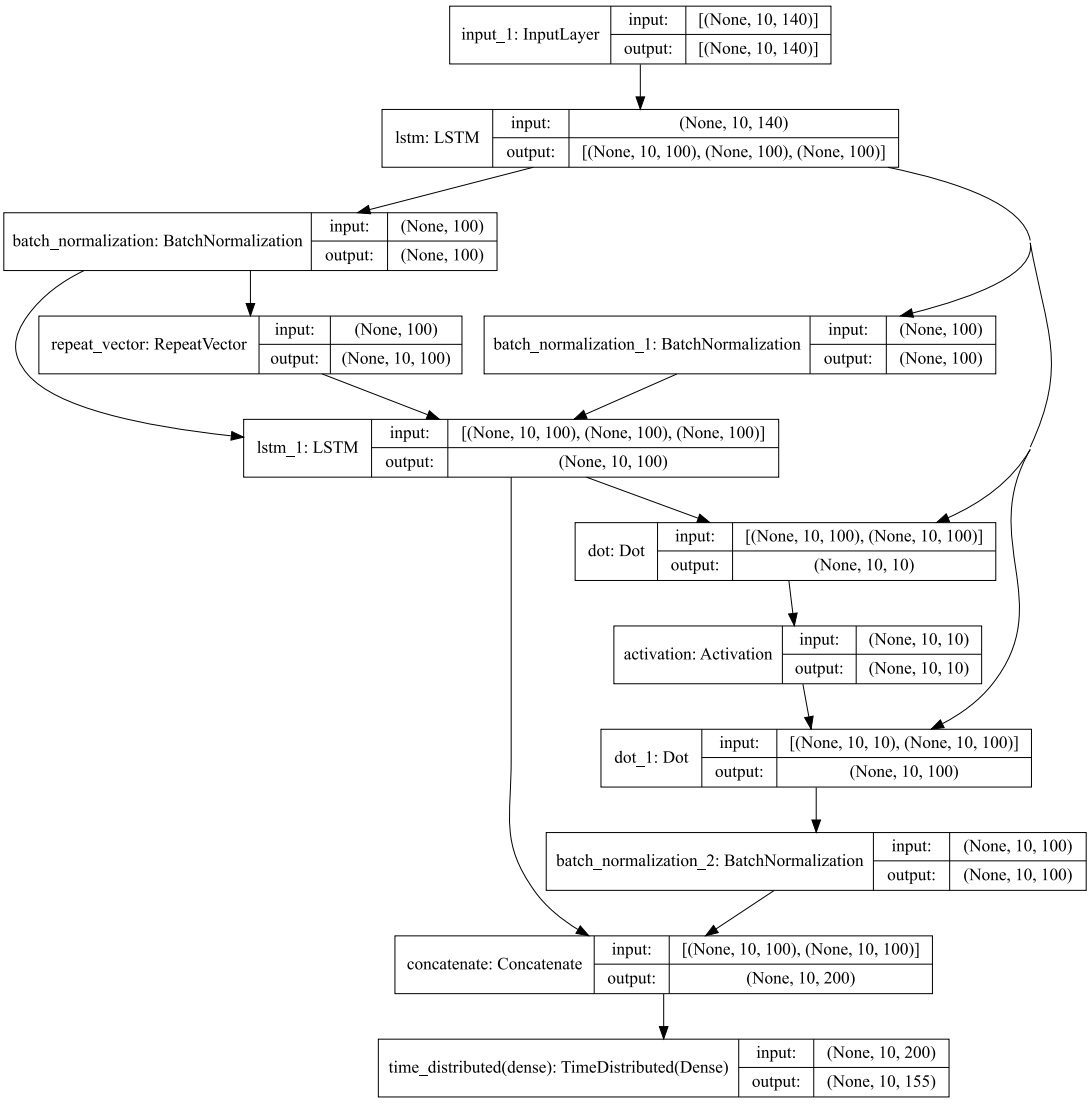


Fig. 8. Schematic diagram of the neural network structure of the Seq2Seq framework (total number of parameters is 2,189,896 and 155 indicates the number of prediction classes).

Table 3
Entity co-occurrence frequency across the dataset. (The horizontal axis represents the socio-economic entity and the vertical axis represents the physical entity).

Frequency	Window	Tree	Building	Car	Skyscraper	Wheel	Street	House	Vehicle	Land
Companies	15422	17659	15007	14783	6548	13151	7064	12677	3837	7076
Street gates	17244	19715	16721	16495	7023	14582	7529	14326	4210	7740
Residential areas	5066	5738	4956	4866	2411	4297	2214	4118	1268	2309
Living service places	17770	20254	17237	16922	7190	15012	7748	14732	4316	7944
Chinese restaurants	16814	19095	16310	16003	6833	14183	7177	13959	4045	7495
Government agencies	2569	3091	2523	2594	1052	2261	1347	2183	642	1314
Parking lots	15264	17433	14858	14629	6595	12897	6725	12548	3749	6834
Agricultural forestry and fishery market	398	606	360	401	175	313	298	323	119	210
Bus stations	10088	11955	9983	9928	4704	8734	5405	8316	2586	5111

extensive urban data for self-supervised learning, we can engage in a "chat" with SVIs and leverage the model to assist in understanding the surrounding environment. This has the potential to bring about revolutionary breakthroughs in the GIS discipline.

Although the method shows promise, some limitations remain, such as using general training data for object detection instead of urban-specific data, and imbalanced entity distributions in the city. In future work, we aim to enhance the feature extraction method by considering the biases of different sampling angles and improving the accuracy of model prediction (Beaucamp et al., 2022).

CRediT authorship contribution statement

Yan Zhang: Conceptualization, Methodology, Writing – original draft. **Fan Zhang:** Writing – review & editing, Supervision. **Libo Fang:** Data curation. **Nengcheng Chen:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, ‘‘Translating built environment to socioeconomic environment–an urban sensing enhancement model".

Data availability

Data will be made available on request.

Appendix

See Fig. 8.

References

Bao, H., Ming, D., Guo, Y., Zhang, K., Zhou, K., Du, S., 2020. DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data. Remote Sens. 12 (7), 1088.

Beaucamp, B., Tourre, V., Servières, M., Leduc, T., 2022. The whole is other than the sum of its parts: Sensibility analysis of 360° urban image splitting. ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci. 4, 33–40.

Biljecki, F., Zhao, T., Liang, X., Hou, Y., 2023. Sensitivity of measuring the urban form and greenery using street-level imagery: A comparative study of approaches and visual perspectives. Int. J. Appl. Earth Obs. Geoinf. 122, 103385.

Boeing, G., 2021. Spatial information and the legibility of urban form: Big data in urban morphology. Int. J. Inf. Manage. 56, 102013.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Cai, B.Y., Li, X., Seiferling, I., Ratti, C., 2018. Treepedia 2.0: applying deep learning for large-scale quantification of urban tree cover. In: 2018 IEEE International Congress on Big Data (BigData Congress). IEEE, pp. 49–56.

Chen, X., Biljecki, F., 2022. Mining real estate ads and property transactions for building and amenity data acquisition. Urban Inform. 1 (1), 12.

Chen, Y., Chen, X., Liu, Z., Li, X., 2020. Understanding the spatial organization of urban functions based on co-location patterns mining: A comparative analysis for 25 Chinese cities. Cities 97, 102563.

Chen, D., Tu, W., Cao, R., Zhang, Y., He, B., Wang, C., Shi, T., Li, Q., 2022. A hierarchical approach for fine-grained urban villages recognition fusing remote and social sensing data. Int. J. Appl. Earth Obs. Geoinf. 106, 102661.

Chen, N., Zhang, Y., Du, W., Li, Y., Chen, M., Zheng, X., 2021. KE-CNN: A new social sensing method for extracting geographical attributes from text semantic features and its application in Wuhan, China. Comput. Environ. Urban Syst. 88, 101629.

Cheng, L., Yuan, Y., Xia, N., Chen, S., Chen, Y., Yang, K., Ma, L., Li, M., 2018. Crowd-sourced pictures geo-localization method based on street view images and 3D reconstruction. ISPRS J. Photogramm. Remote Sens. 141, 72–85.

Deng, M., Yang, W., Chen, C., Wu, Z., Liu, Y., Xiang, C., 2021. Street-level solar radiation mapping and patterns profiling using baidu street view images. Sustainable Cities Soc. 103289.

Dimakis, A.D., Sarwate, A.D., Wainwright, M.J., 2008. Geographic gossip: Efficient averaging for sensor networks. IEEE Trans. Signal Process. 56 (3), 1205–1216.

Du, Z., Zhang, X., Li, W., Zhang, F., Liu, R., 2020. A multi-modal transportation data-driven approach to identify urban functional zones: An exploration based on Hangzhou city, China. Trans. GIS 24 (1), 123–141.

Fan, C., Jiang, Y., Mostafavi, A., 2020. Social sensing in disaster city digital twin: Integrated textual–visual–geo framework for situational awareness during built environment disruptions. J. Manage. Eng. 36 (3), 04020002.

Feng, Y., Tong, X., 2020. A new cellular automata framework of urban growth modeling by incorporating statistical and heuristic methods. Int. J. Geogr. Inf. Sci. 34 (1), 74–97.

Gao, S., Janowicz, K., Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans. GIS 21 (3), 446–467.

Huang, L., Ma, Y., Wang, S., Liu, Y., 2019. An attention-based spatiotemporal lstm network for next poi recommendation. IEEE Trans. Serv. Comput..

Huang, H., Yao, X.A., Krisp, J.M., Jiang, B., 2021. Analytics of location-based big data for smart cities: Opportunities, challenges, and future directions. Comput. Environ. Urban Syst. 90, 101712.

Keralis, J.M., Javanmardi, M., Khanna, S., Dwivedi, P., Huang, D., Tasdizen, T., Nguyen, Q.C., 2020. Health and the built environment in United States cities: Measuring associations using google street view-derived indicators of the built environment. BMC Public Health 20 (1), 1–10.

Khosla, A., An An, B., Lim, J.J., Torralba, A., 2014. Looking beyond the visible scene. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3710–3717.

Kim, W., Son, B., Kim, I., 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 139, PMLR, pp. 5583–5594, URL http://proceedings.mlr.press/v139/kim21k.html.

Li, S., Dragicevic, S., Castro, F.A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., et al., 2016. Geospatial big data handling theory and methods: A review and research challenges. ISPRS J. Photogramm. Remote Sens. 115, 119–133.

Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81.

Liu, Q., Gu, J., Yang, J., Li, Y., Sha, D., Xu, M., Shams, I., Yu, M., Yang, C., 2021. Cloud, edge, and mobile computing for smart cities. In: Urban Informatics. Springer, pp. 757–795.

Liu, K., Yin, L., Lu, F., Mou, N., 2020. Visualizing and exploring POI configurations of urban regions on POI-type semantic space. Cities 99, 102610.

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., et al., 2023. On the opportunities and challenges of foundation models for geospatial artificial intelligence. arXiv preprint arXiv:2304.06798.

Mohamed, S.A., Elsayed, A.A., Hassan, Y., Abdou, M.A., 2021. Neural machine translation: past, present, and future. Neural Comput. Appl. 1–13.

O’Halloran, K.L., 2011. Multimodal discourse analysis. Companion Discourse. London New York: Continuum.

Qian, X., Wu, Y., Li, M., Ren, Y., Jiang, S., Li, Z., 2020. Last: Location-appearance-semantic-temporal clustering based POI summarization. IEEE Trans. Multimed. 23, 378–390.

Qin, K., Xu, Y., Kang, C., Kwan, M.P., 2020. A graph convolutional network model for evaluating potential congestion spots based on local urban built environments. Trans. GIS 24 (5), 1382–1401.

Salim, F.D., Dong, B., Ouf, M., Wang, Q., Pigliautile, I., Kang, X., Hong, T., Wu, W., Liu, Y., Rumi, S.K., et al., 2020. Modelling urban-scale occupant behaviour, mobility, and energy in buildings: A survey. Build. Environ. 183, 106964.

Shi, K., Wang, Y., Lu, H., Zhu, Y., Niu, Z., 2021. EKGTF: A knowledge-enhanced model for optimizing social network-based meteorological briefings. Inf. Process. Manage. 58 (4), 102564.

Sievert, C., Shirley, K., 2014. LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. pp. 63–70.

Sotomayor-Gómez, B., Samaniego, H., 2020. City limits in the age of smartphones and urban scaling. Comput. Environ. Urban Syst. 79, 101423.

Stiles, J., Li, Y., Miller, H.J., 2022. How does street space influence crash frequency? An analysis using segmented street view imagery. Environ. Plan. B: Urban Anal. City Sci. 23998083221090962.

Stock, K., 2018. Mining location from social media: A systematic review. Comput. Environ. Urban Syst. 71, 209–240.

Tingting, Z., Wei, T., Yang, Y., Chen, Z., Tianhong, Z., Qiuping, L., Qingquan, L., 2020. Sensing urban vibrancy using geo-tagged data. Acta Geod. Cartogr. Sinica 49 (3), 365.

Tu, W., Cao, J., Yue, Y., Shaw, S.L., Zhou, M., Wang, Z., Chang, X., Xu, Y., Li, Q., 2017. Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. Int. J. Geogr. Inf. Sci. 31 (12), 2331–2358.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008.

Wang, P., Hu, T., Gao, F., Wu, R., Guo, W., Zhu, X., 2022a. A hybrid data-driven framework for spatiotemporal traffic flow data imputation. IEEE Internet Things J. 9 (17), 16343–16352.

Wang, P., Zhang, Y., Hu, T., Zhang, T., 2022b. Urban traffic flow prediction: A dynamic temporal graph network considering missing values. Int. J. Geogr. Inf. Sci. 1–28. http://dx.doi.org/10.1080/13658816.2022.2146120.

Wang, P., Zhang, T., Zheng, Y., Hu, T., 2022c. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. Int. J. Geogr. Inf. Sci. 36 (6), 1231–1257.

Xu, L., Chen, N., Chen, Z., Zhang, C., Yu, H., 2021. Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions. Earth-Sci. Rev. 103828.

Xu, L., Du, Z., Mao, R., Zhang, F., Liu, R., 2020. GSAM: A deep neural network model for extracting computational representations of Chinese addresses fused with geospatial feature. Comput. Environ. Urban Syst. 81, 101473.

Xue, B., Xiao, X., Li, J., 2020. Identification method and empirical study of urban industrial spatial relationship based on POI big data: A case of Shenyang city, China. Geogr. Sustain. 1 (2), 152–162.

Yang, C., Zeng, W., Yang, X., 2020. Coupling coordination evaluation and sustainable development pattern of geo-ecological environment and urbanization in Chongqing municipality, China. Sustainable Cities Soc. 61, 102271.

Yin, J., Dong, J., Hamm, N.A., Li, Z., Wang, J., Xing, H., Fu, P., 2021. Integrating remote sensing and geospatial big data for urban land use mapping: A review. Int. J. Appl. Earth Obs. Geoinf. 103, 102514.

Yue, Y., Zhuang, Y., Yeh, A.G., Xie, J.Y., Ma, C.L., Li, Q.Q., 2017. Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy. Int. J. Geogr. Inf. Sci. 31 (4), 658–675.

Zhang, Y., Chen, N., Du, W., Li, Y., Zheng, X., 2021a. Multi-source sensor based urban habitat and resident health sensing: A case study of Wuhan, China. Build. Environ. 198, 107883.

Zhang, Y., Chen, Z., Zheng, X., Chen, N., Wang, Y., 2021b. Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data. J. Hydrol. 603, 127053.

Zhang, X., Du, S., Wang, Q., 2017. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. ISPRS J. Photogramm. Remote Sens. 132, 170–184.

Zhang, F., Fan, Z., Kang, Y., Hu, Y., Ratti, C., 2021c. "Perception bias": Deciphering a mismatch between urban crime and perception of safety. Landsc. Urban Plan. 207, 104003.

Zhang, Y., Liu, P., Biljecki, F., 2023. 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., 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, Y., Zhang, F., Chen, N., 2022a. 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., 2018. Representing place locales using scene elements. Comput. Environ. Urban Syst. 71, 153–164.

Zhang, J., Zhang, L., Qin, Y., Wang, X., Zheng, Z., 2020a. Influence of the built environment on urban residential low-carbon cognition in zhengzhou, China. J. Clean. Prod. 271, 122429.

Zhang, Y., Zheng, X., Helbich, M., Chen, N., Chen, Z., 2022b. City2vec: Urban knowledge discovery based on population mobile network. Sustainable Cities Soc. 85, 104000.

Zhang, F., Zu, J., Hu, M., Zhu, D., Kang, Y., Gao, S., Zhang, Y., Huang, Z., 2020b. Uncovering inconspicuous places using social media check-ins and street view images. Comput. Environ. Urban Syst. 81, 101478.

Zhao, T., Liang, X., Tu, W., Huang, Z., Biljecki, F., 2023. Sensing urban soundscapes from street view imagery. Comput. Environ. Urban Syst. 99, 101915.

Zheng, X., Han, J., Sun, A., 2018. A survey of location prediction on twitter. IEEE Trans. Knowl. Data Eng. 30 (9), 1652–1671.

Zhou, B., Zou, L., Mostafavi, A., Lin, B., Yang, M., Gharaibeh, N., Cai, H., Abedin, J., Mandal, D., 2022. VictimFinder: Harvesting rescue requests in disaster response from social media with BERT. Comput. Environ. Urban Syst. 95, 101824.

Zhu, Y., Zhu, A.X., Feng, M., Song, J., Zhao, H., Yang, J., Zhang, Q., Sun, K., Zhang, J., Yao, L., 2017a. A similarity-based automatic data recommendation approach for geographic models. Int. J. Geogr. Inf. Sci. 31 (7), 1403–1424.

Zhu, Y., Zhu, A.X., Song, J., Yang, J., Feng, M., Sun, K., Zhang, J., Hou, Z., Zhao, H., 2017b. Multidimensional and quantitative interlinking approach for linked geospatial data. Int. J. Digit. Earth 10 (9), 923–943.