This is the Accepted Manuscript version of an article published by Elsevier in the journal *ISPRS Journal of Photogrammetry and Remote Sensing* in 2023, which is available at: https://doi.org/10.1016/j.isprsjprs.2023.03.008
Cite as: 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 Journal of Photogrammetry and Remote Sensing*, 198: 153–168.

Knowledge and Topology: A Two Layer Spatially Dependent Graph Neural Networks to Identify Urban Functions with Time-series Street View Image

Yan Zhang[a,b], Pengyuan Liu[b] and Filip Biljecki[b,c,*]

[a]State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China
[b]Department of Architecture, National University of Singapore, Singapore
[c]Department of Real Estate, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
GeoAI
Natural Language Processing
GeoKG
Pretrained Model
Knowledge Graph
Multi-Modal

ABSTRACT

With the rise of GeoAI research, streetscape imagery has received extensive attention due to its comprehensive coverage, abundant information, and accessibility. However, obtaining a holistic spatial-temporal scene representation is difficult because places are often composed of multiple images from different angles, times and locations. This problem also exists in other types of geo-tagged imagery. To solve it, we propose a purely visual, robust, and reliable method for urban function identification at the street scale. We introduce a method based on a two-layer spatially dependent graph neural network structure, which handles sequential street view imagery as input (typically available in services such as Google Street View, Baidu Maps, and Mapillary), with full consideration of the spatial dependencies among road networks. In this paper, we construct an urban topological map network using OpenStreetMap data in Wuhan, China, and compute a semantic representation of the scene as a whole at the street scale using a large-scale pre-trained model. We construct the graph network with streets as nodes based on 28,693 mapping relationships constructed from 75,628 street view images and 5,458 streets. Only 5.3% of the node labels were required to obtain 10 categories of functions for all nodes in the study area. The results demonstrate that by using appropriate spatial weights, street encoder, and graph structure, our novel method achieves high accuracy of P@1 46.2%, P@3 73.0%, P@5 82.4%, and P@10 89.9%, fully demonstrating the effectiveness of the introduced approach. We also use the model to sense urban spatial-temporal renewal by computing time series street images. The model is also applicable to the prediction of other attributes, where only a small number of labels are required to obtain valid and reliable scene perception results. The example data and code is shared at: github.com/yemanzhongting/Knowledge-and-Topology.

1. Introduction

Urban functions are aggregated areas resulting from various types of human activities in urban space (Zhou et al., 2020; Crooks et al., 2015). As the continuous expansion and development of cities lead to an increasing need of monitoring and updating urban functions, we have witnessed numerous efforts devoted into such studies (Hu et al., 2021; Lu et al., 2022; Biljecki and Chow, 2022) in the recent years. The city is the most dominant carrier of human activities (Liu et al., 2018). Thanks to physical and social sensing networks, and the proliferation of big data technologies, a wealth of multi-modal urban datasets are generated, which has brought massive changes in the study of urban functions (Zhang et al., 2019a).

Remote sensing imagery is an important data source for data-driven urban function studies, providing a bird's eye view that has a successful history of capturing an overall understanding on land use (e.g., lakes, farmland, buildings, etc.) (Cao et al., 2018, 2020). However, such imagery-based analysis lacks in-situ socio-economic semantic information for understanding human-space interactions further (Zhang et al., 2021c). Thus, the quality of urban function identification using such top view imagery is often inadequate (Van de Voorde et al., 2011). Meanwhile, the proximate sensing perspectives such as streetscape (or street view) imagery (Qiao and Yuan, 2021) captured at high spatial

*Corresponding author
sggzhang@whu.edu.cn (Y. Zhang); pyliu93@nus.edu.sg (P. Liu); filip@nus.edu.sg (F. Biljecki)
ORCID(s):

resolution facilitate obtaining detailed urban information (Zhang et al., 2020; Chen et al., 2022). Numerous existing studies have asserted that street view images are efficient in identifying urban functions (Xu et al., 2022b; Biljecki and Ito, 2021), sometimes in conjunction with other data sources such as point of interest (POI) (Hu et al., 2020b; Yao et al., 2017), social media (Chen et al., 2017; Gao et al., 2017), taxi trajectory (Hu et al., 2021), and night light remote sensing (Huang et al., 2021).

According to Liu et al. (2021), 51% of the urban function identification study used only a single data source, 49% used two or more of the above datasets (Liu et al., 2017; Yin et al., 2021), and POI serves as the most widely used data, being used by 75% of studies. However, in existing research, imagery of the streetscape has only been considered a secondary data source, complementing other data sources (e.g. remote sensing), to improve the identification accuracy (Fang et al., 2021; Qiao and Yuan, 2021). This gap leads to devise the first research question of whether we can perceive urban functions solely from visual data.

Taking advantage of the rapid progress of digital image processing, many studies used neural networks for image classification (Hu et al., 2020a; Kang et al., 2018), semantic segmentation (Lauko et al., 2020; Qi et al., 2020) and object detection (Campbell et al., 2019; Chen et al., 2020) the street scenes in the images. These studies, however, can only address a single streetscape image. They cannot be generalized to a broader area (Zhang et al., 2021c), as they cannot provide an effective holistic representation of spatial units, such as for Area of Interest (AOI) (Li et al., 2021a), Traffic Analysis Zone (TAZ) (Gong et al., 2020; Chen et al., 2021a) or building footprints (Song et al., 2022). Generating an overall semantic representation of the space is the second question answered in this paper.

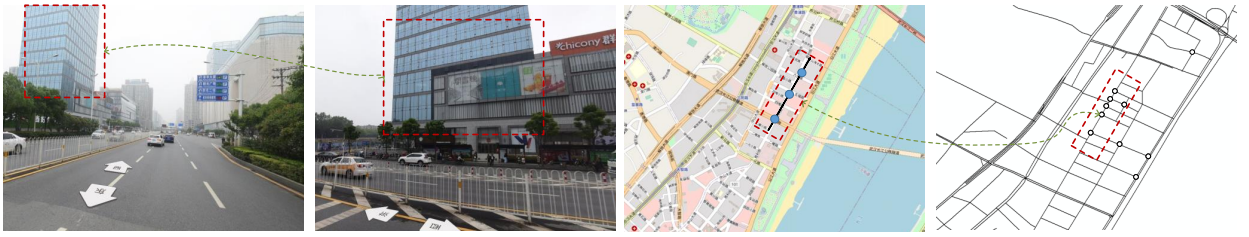
For sequential input data sources such as street view images, which are common in both commercial and crowd-sourced platforms (e.g. Google Street View, Baidu Maps, Tencent Maps, Mapillary, and KartaView), it is essential to note that the pre-trained convolutional neural network (CNN)-based methods (e.g., Resnet) can only process images one by one (Zhang et al., 2021b; Yao et al., 2021). The typical way to obtain a spatial unit embedding is to calculate the average of the multiple images feature vectors. However, such an approach ignores the location information of the imagery and prevents the model from capturing the spatial heterogeneity of urban functional areas. That is, it will not capture the spatial topological relationships between those images, and also, the spatial location relationships between key geographic entities within the images (Fang et al., 2021). Moreover, there is information bias in representing spatial units with separate or small number of images (Wang et al., 2021; Kang et al., 2018), and the loss of spatial contextual information can lead to different urban function identifications for the same region (Amiruzzaman et al., 2021).

As illustrated in Figure 1, the results of both semantic segmentation and object detection performed on street view imagery lack comprehensive semantic information. Even if two images have similar proportions of various elements after semantic segmentation, or the same objective features exist after object detection, there may still be considerable geographical differences in the actual situation described by the two images due to the different spatial relationships between the elements or features. More specifically, the street view images in Figure 1 both contain entities such as Building, Window, Tree, Street Light, Motorcycle, and Car. Besides, the pixel proportions occupied by these entities in the two images are similar. For example, the pixel ratio of the building on the Figure 1 left image is 23%, the ratio of green plants is 20%, and the ratio of pixels occupied by the sky is 28%, while the pixel ratio of the building on the Figure 1 right image is 26%, the green view index is 28%, and the sky view factor is also 28%. With the above information, it is difficult to find the difference between the street view images of two locations simply by the type of entity, the number of entities, or the pixel ratio of the field of view.

To capture the spatial relationships between the street scenes, such as the ones exhibited in the Figure 2, we use road data as a network structure and choose the street with multiple streetscapes as the minimum spatial unit. We design to capture two layers of spatial relationship, the first one is to capture overlapping entities in different views, the second one is to capture the spatial topology relationship between streets. As the data source, we rely on OpenStreetMap (OSM) thanks to its global coverage, ease of access, and quality evidenced by many urban studies (Chen et al., 2021b; Pađen et al., 2022; Venerandi et al., 2022), but our approach is generalisable to other data sources.

Graph Neural Networks (GNN) can handle non-Euclidean structure data and extract spatial features from the topological graph for learning efficiently (Wang et al., 2022b,a). Existing studies have proved that the performance of GeoAI models can be improved by considering the spatial topology (Zhu et al., 2020). Besides, GNNs are able to learn features of the nodes based on their local neighborhood, which is defined by the edges connecting them. This allows GNNs to work with only a small subset of the labelled nodes, or even with completely unlabeled graphs (Zhao et al., 2022). While urban function (land use) identification is a very typical label-scarcity type task, spending a lot of manpower on land use labeling is time-consuming and labor-intensive, and semi-supervised methods like GNN are

Figure 1: Examples of object detection and semantic segmentation on two scenes obtained from street view imagery. Although the analysis results of these two street views are similar, there are considerable geographical differences in the actual conditions. The first was taken on a school road and the second on a commercial street, pieces of information that may be instrumental for urban studies but out of reach of conventional methods using street-level imagery. Thus, we posit that traditionally used approaches do not give full justice to the urban function identification and their distinction, and propose an enhanced approach relying on the sequential nature of imaging and a graph neural network structure. Source of the imagery: Tencent Street View.



Layer 1: Capture overlapping entities

Layer 2: Capture spatial topology

Figure 2: Schematic diagram of two-layers spatial relationship capture. For example, we identify a building from one image, and the same building may also appear in other images, which establishes a relationship between the two images. Source of the imagery: Tencent Street View. Source of the map: © OSM contributors.

very useful for such task (Zhang et al., 2022a). In view of the above advantages, we have witnessed an increasing use of GNNs in urban-related studies (e.g. road traffic forecasting) (Zhao et al., 2019; Yu et al., 2020; Liu and Biljecki, 2022; Abdelrahman and Miller, 2022).

Considering that urban functions will have certain co-location patterns or spatial dependencies (Yu et al., 2017), we introduce GNN to capture the topological relationships of roads in order to improve recognition accuracy. Related research has attracted considerable attention recently. For example, Hu et al. (2021) used traffic trajectory data, treated roads as nodes and extracted the topological relations between them. Nonetheless, such a relationship lacks semantic information and the description of physical environment. This limitation is why the trained model is not rich in predictions and can only identify three types of functions: commercial, public, and traffic. Xu et al. (2022c) described city region features based on POI data and relative relationship between POIs, extended the prediction results to six categories. However, it can only use the category attributes of the POI (a standard treatment is to analogize POI categories, such as business, traffic, etc. to 'words', and regions to 'documents', but much information is lost in such methods (Yao et al., 2017)), and lacks the use of road network topology information (Inoue et al., 2022).

Inspired by the application of vision-language multi-modality in remote sensing image analysis (Wang et al., 2022b), we propose a two-layer spatially dependent graph neural network based on knowledge (socio-economic information extracted from the physical environment of the streetscape) — topology (neighborhood relationships between road networks). Using the urban road networks as a backbone, we generated a street semantic knowledge graph and computed an embedding representation for each street node. The first spatially dependent layer is a knowledge layer that generates objective descriptions of street view images in 'human language', which are then aggregated to the whole street. In other words, calculating the caption of street scene. Machine reasoning is used to deconstruct and understand the content of the image, then generate a natural semantic description of the given scene (Hossain et al., 2019).

To the best of our knowledge, some attempts have been proved to be successful using such a cross-modal approach in understanding remote sensing images (Shen et al., 2020; Li et al., 2021c), and such studies are highly valuable for applications in disaster assessment, urban planning and geographic information retrieval (Hong et al., 2020; Murali and Shanthi, 2022). The cross-modal model has the ability to describe the objects, attributes and relationships between objects that appear in the image (Yang et al., 2022b), which will contain more information compared to the traditional results of semantic segmentation (Amiruzzaman et al., 2021) and object detection (Ning et al., 2022). Accurate scene inference is a challenging task that requires a fine-grained understanding of global and local entities in an image, as well as their attributes and relationships, and also requires a joint contribution from the fields of computer vision and natural language processing (Zhang et al., 2022a).

In this paper, we refer to the knowledge graph consisting of geographical entities and spatial location relationships between entities in the streetscape as the street semantic knowledge graph (von Richthofen et al., 2022; Li et al., 2021b; Chadzynski et al., 2022). The cross-modal technique is applied to urban street view interpretation to generate the scene caption and the entity-relationship-entity knowledge triple.

The second spatially dependent layer is the topology layer, which constructs a spatial weight matrix based on the road spatial topology. It can take full advantage of the spatial dependence of the urban functions distribution (Georganos et al., 2021). Our method can fully use the graph semi-supervised learning features to obtain accurate prediction results with only a small number of training labels. Considering that in a spatial unit, street view images are often sampled in different times, this means that not only socio-economic and physical built information can be perceived, but also temporal-spatial changes, which was often neglected (Xu et al., 2022a; Biljecki and Ito, 2021).

The main contributions of this paper are threefold:

- The first is a method for sensing urban functions based on pure visual data, which is compatible with arbitrary images containing geo-tags and has significantly broadened the application scenarios of proximate sensing images in this field.
- The second is the solution to the sequence input and regional representation of street view images by effectively using their spatial location information. It can perceive city temporal-spatial changes (city renewal) and provide more accurate street function prediction products.
- The third is the generation of the street semantic knowledge graph based on the intermediate products of the computation (physical entities of the city, spatial location relationships of the entities) to improve the urban comprehensibility.

On a broader scope, we also introduce a new use case of street view imagery, a rapidly growing source of urban data, but not utilised for this purpose hitherto.

Our study is organized as follows: the second section presents the model architecture of our Spatio-temporal two-layer graph neural network; the third section is about the experiments, which discusses the model accuracy and the Spatio-temporal representation performance, then generates the knowledge graph for visualization; the last section summarizes this work and discusses the advantages and applicability of the model, which is of high value for urban planning and geographic information retrieval, etc.

2. Method

Our method is divided into five parts, as shown in Figure 3, Step 1 and Step 2 are the encoder part of the model, and Step 3, Step 4, Step 5 are the decoder part of the model.

Step 1: Cross-modal extraction scenario description (Street View Captioning)

Each source or form of information can be referred to as a modality. Cross-Modal Machine Learning (CMML) aims to achieve the ability to process and understand data from multiple sources of modality through a machine learning approach (Lin et al., 2016). The descriptive text of images (caption) is a cross-modal process that converts images into textual descriptions and provides rich semantic information for further computation. There are three methods to generate image captions: template-based, retrieval-based, and sequence-generation-based method (Zhao, 2021). The last method can not only obtain the correspondence between image features and words but also learn the sequence relationship between adjacent words and generate flexible and variable descriptions, which is the method adopted in this paper.

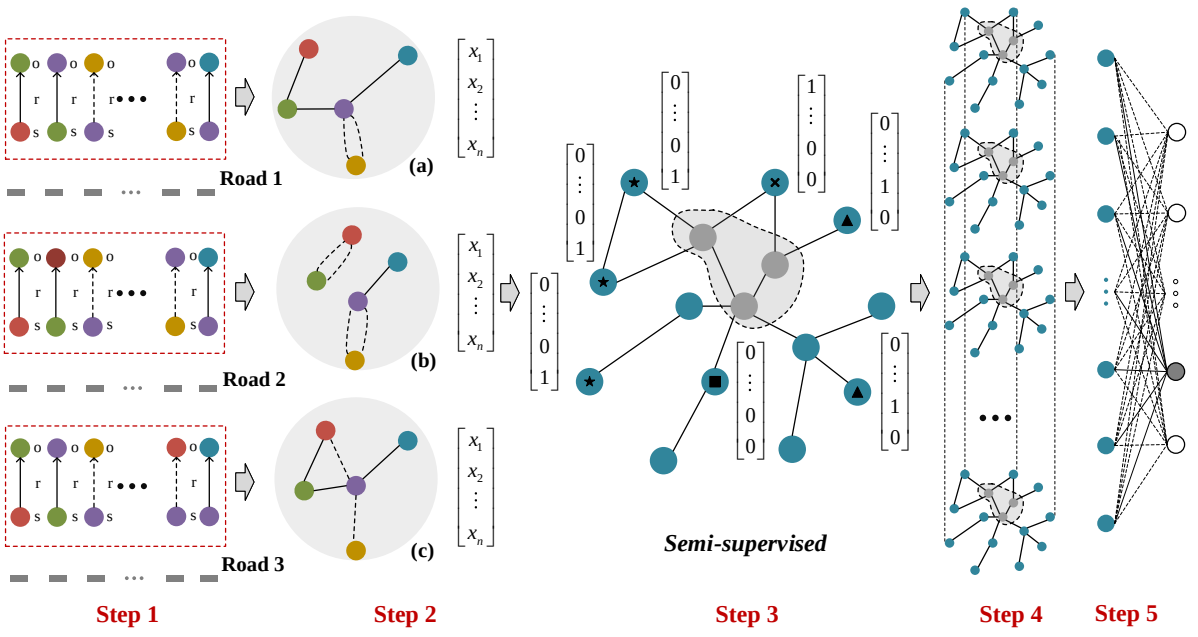


Figure 3: Schematic diagram of the Knowledge-Topology Two Layers Spatially Dependent Graph Neural Network.

Figure 4 shows the detailed steps of the cross-modal computation, where we input a street scene image and output the corresponding text description, "A red car parked on the side of the road". The description contains the vital physical entities of the city, the nature of the entities, and the spatial location relationship between the entities. As such, we can extract a triple like ("A red car" ->s," parked on the side of "->r," the road" ->o). We perform a similar calculation for each image, aggregating it to form a semantic description of each street.

We use a vision-language model (Figure 4) trained through the Bottom-Up and Top-Down Attention to obtain urban street view captions (Anderson et al., 2018; Rennie et al., 2017; Lu et al., 2020), its applicability in street views can be found in (Zhang et al., 2022a). During the pre-training process, the model's input consists of images and the corresponding description text. Note that there is no relevant caption dataset in the street view domain. Therefore, the model used the Microsoft COCO caption dataset (Lin et al., 2014; Lu et al., 2017) for the model training step, which consists of a variety of scenarios, both indoor and outdoor.

The sequence generation unit consists of two LSTM layers (Hochreiter and Schmidhuber, 1997) and an attention layer. As shown in Figure 4, the model uses an object detection neural network (Faster R-CNN (Ren et al., 2015)) to extract the image features and divides the image into k regions, which are fed into the recurrent neural network together with a text description containing N words (tokens).

In Figure 4, W_e is the word vector matrix, $\prod_t$ is the one-hot encoding of the word at time t, and the product of the two represents the word vector of the input word y_t at that moment. $\bar{v}$ denotes the mean-pooled of an image feature, v_i means the image feature of the ith region. Then the conditional distribution over possible output words at the time step t is:

$$p\left(y_t|y_{1:t-1}\right) = softmax\left(W_p h_t^2 + b_p\right) \tag{1}$$

Where W_p is the parameter we need to train, b_p is the weights and biases to be learned, and the output y is the word, i.e., the textual description of the street image.

Step 2: Aggregate the images and calculate the scene embedding of the street

Based on the scene description generated by Step 1, we treat the street as a 'document' containing multiple 'sentences' (street scenes). Through the mapping between street scenes and road networks, we can obtain street-view neighborhood information, then calculate the feature code of the street as a whole.

We encode the 'document' using the Bidirectional Encoder Representations from Transformers (BERT) model

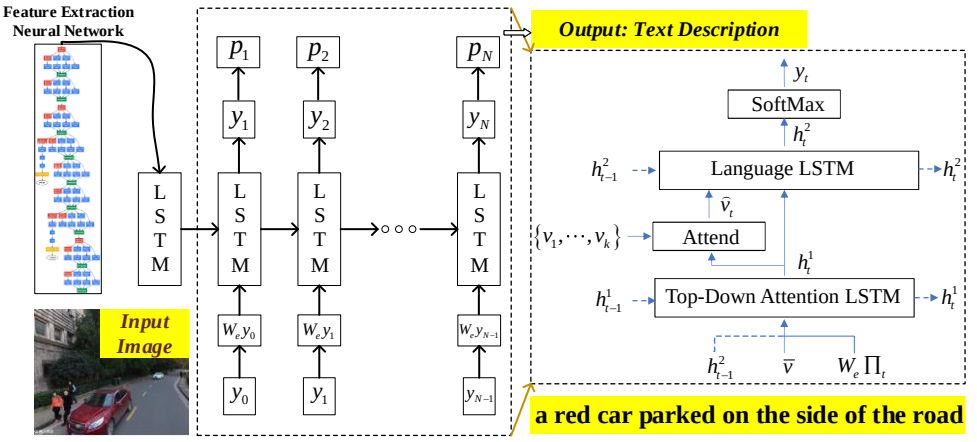


Figure 4: Obtain street view descriptions based on the vision-language pre-trained model. The bottom-up mechanism (based on Faster R-CNN) extracts image regions and corresponding feature vector, the top-down mechanism adjusts feature weighting. Source of the imagery: Tencent Street View.

Table 1
Semantic Embedding of Street Scenes based on Bert Encoder

Dim \ Street	0	1	2	3	4	5	6	7	8	...	5457
0	-0.115	-0.436	-0.220	-0.464	-0.444	-0.553	-0.637	-0.397	-0.368	...	-0.650
1	-0.693	0.281	-0.226	-0.729	-0.263	-0.168	0.243	0.092	0.017	...	0.285
2	-0.791	-0.359	-0.100	-0.690	-0.410	-0.188	-0.404	-0.249	-0.374	...	-0.280
3	0.359	-0.072	0.086	0.153	0.204	0.107	0.038	-0.017	-0.119	...	-0.022
4	0.070	-0.460	-0.218	-0.048	-0.266	-0.019	-0.341	-0.300	-0.190	...	-0.358
5	0.282	-0.613	-0.466	-0.389	-0.343	-0.796	-0.586	-0.580	-0.469	...	-0.570
6	0.099	-0.745	-0.691	-1.046	-0.657	-1.065	-0.523	-0.450	-0.888	...	-0.523
7	-0.249	0.492	0.754	0.098	0.435	0.322	0.300	0.523	0.489	...	0.458
...	...	...	...	...	...	...	...	...	...	...	
767	0.087	-0.190	-0.107	-0.216	-0.225	-0.125	-0.140	-0.059	-0.383	...	-0.067

(Vaswani et al., 2017), which consists of 12 transformer layers, 12 attention heads, 768 dimensions, and 110M parameters. Subsequently, a $D(768)$ dimensional scene embedding is generated as the model input for initialization. In addition to Bert, the struct in Figure 3 can also be compatible with various embedding methods, such as Doc2vec (Niu and Silva, 2021). The experimental section will also compare the computational effects of different embedding methods. Table 1 represents the results of the street embedding calculation with Bert as the semantic encoder (5458 streets, 768 dimensions).

Step 3: Labelling and spatial weights (road network topology)

The streetscape is rich in socio-economic information (Zhang et al., 2021a, 2018), and this paper assesses its ability to identify urban street functions. We use the EULUC-China dataset produced by (Gong et al., 2020) to label the streets. The EULUC-China dataset identifies 27 major cities in China based on multiple data sources such as 10 m satellite images, OpenStreetMap, night time lights, POIs, and Tencent Social Big Data, including five primary classification labels and 12 secondary classification labels, with an overall classification accuracy of 60%. The urban land use classification labels refer to the Chinese land-use status classification standard (GB/T 21010-2017). Among those 12 secondary categories, road classification is not the type we focused, and the study area does not contain Airport facilities type, we excluded such irrelevant labels. As a result, we constructed a street function dataset containing ten types of labels in Table 2, detailed information about labels can be found in the literature (Gong et al., 2020).

The study area contains 13,889 streetscape sampling locations, and each has four images from different angles,

Table 2
Categories, descriptions, and labeling status of urban functions.

ID	Type	Description	Region Number	Labeled Street
0	Residential	Houses and apartment where people live	1,096	87
1	Business office	Commercial office space	114	15
2	Commercial service	Commercial retails,restaurants,lodging,and entertainments	161	29
3	Industrial	Manufacturing,warehouse,mining,etc.	176	37
4	Transportation stations	Transportation facilities	24	6
5	Administrative	Government,public service agencies	61	6
6	Educational	Education and research	289	65
7	Medical	Hospitals	28	12
8	Sport and cultural	Public sports and training,cultural services	71	28
9	Park and greenspace	Entertainments and environmental conservations	273	10

Note: There are 5,458 streets, 13,889 sampling points and 2,296 regions, and only labeled about 5.3% of the streets.

5,458 OSM roads, and 2,296 functional regions with labels. To obtain the labels, we set a 25 m wide buffer for the sampled locations (50m can cover most of the city road width), which is spatially connected to the functional regions to label the street view image. As elaborated in Step 1, we have constructed the relationship between streetscape and street, therefore, here, we simply count the streetscape labels that contain the most categories as the street's primary function. Some streets contain multiple labeled street views. For those streets, we select streets with more prominent functions as training data. As shown in Table 2. We labeled 190 streets, with an overall annotation rate of 5.3%.

Using the road network topology, we generated a spatial weight matrix to generate the adjacent road relationship. Formally, we used an $n \times n$ (n is the number of streets) matrix A to express it, if there is an adjacent relationship between streets i and j, then $a_{i,j}$ is assigned to 1, otherwise $a_{i,j}$ is assigned to 0.

We used two methods to express the adjacent relationship. The first is Queen contiguity spatial weight method (Suryowati et al., 2018). If there is an intersection and overlap between streets, the technique will mark that the two streets are adjacent. The advantage is that it can better reflect the existing road network, but the disadvantage is that it cannot handle streets without neighbors. The other is K-nearest neighbors spatial weight method. It can calculate the K nearest streets and mark them adjacent relationships. The advantage is that it avoids dangles roads. The disadvantage is that some non-intersecting streets will also be marked adjacent (Zhu et al., 2020).

Step 4: Build semi-supervised graph neural network for training

There are many graph models, and most graph models share their filter parameters over all locations in the graph. Those models are Graph Convolutional Networks (GCN) (Kipf and Welling, 2017; Zhang et al., 2019b). GCNs have an excellent ability to extract graph features. Thus, suitable for semi-supervised learning tasks (only a small amount of labels required), and it takes fewer iterations to converge (Bruna et al., 2014; Kipf and Welling, 2017). We generated a city road graph according to the adjacency matrix A built-in Step 3. Every street was regarded as a node, and our task was to label street functions. This graph includes N nodes, with D dimensions, as the initial feature matrix X. F is the output dimension of the model, then the output is a matrix Z of size $D \times F$, and H represents the intermediate state of the model. The GCN propagation rules are defined as follows:

$$f(H^{(l)}, A) = \sigma\left(AH^{(l)}W^{(l)}\right) \tag{2}$$

Where $W^{(l)}$ is a weight matrix for the lth network layer,$\sigma(\cdot)$ is non-linear activation function $ReLU$. There are still two limitations of the propagation rule at this point. The first is that A is the adjacency matrix of the graph, and the nodes' features are not considered; the second is that the regularization of A is also required. We define $\hat{A} = A + I$, I as the identity matrix. $\hat{D}$ is the diagonal node degree matrix of $\hat{A}$. At this point, the propagation rule of the GCNs is defined as:

$$f(H^{(l)}, A) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{3}$$

The model training procedure is defined as follows: 1. perform forward propagation of the model through the f function. 2. Compute the cross-entropy loss on the labels of known nodes.

$$cross\_entropy = -\sum_{k=1}^{N} \left(p_k \times \log q_k\right) \tag{4}$$

p is the true label, q is the predict label. 3. Back propagate the loss and update each layer's weight matrix W.

The number of layers is the farthest distance a node feature can reach. In a 1-layer GCN, each node can only get information from its neighbors, while in a 2-layer GCN, each node can also get information from its neighbors' neighbors. More layers are not always better; over-stacked layers will cause the model to lose the ability to extract local features. Xu et al. (2022c) believes that in the urban task, the model accuracy decreases as the depth increases, and better accuracy can be achieved by taking two neural network layers on average. Collecting information is carried out independently, with all nodes performing it at the same time (Yang et al., 2022a). At last, the features of the graph are reduced from an initial 768 dimensions to a final ten dimensions. After the cross-entropy loss function (equation 4) training the weight matrix W, and the propagation rule equation 3, we output the prediction result (matrix Z, equation 5) based on the softmax function, i.e., the probability that each node belongs to the above 10 classes.

$$Z = f(X, A) = soft\max\left(\hat{A}\,\mathrm{ReLU}\left(\hat{A}XW^{(0)}\right)W^{(1)}\right) \tag{5}$$

We used three semi-supervised graph neural networks to test the performance, the first one is traditional GCN, which incorporates the whole graph into the calculation, and the relationship between nodes is equal weight. We designed two types of structures, the hidden channel (computational units) is 64 and 32, 32 and 16, respectively. The second one is graph attention network (GAT) (Veličković et al., 2017), which is different from GCN, the relationships between nodes are unequal-weighted, and these weight parameters are also the object of model learning. The third one is GraphSage (Hamilton et al., 2017), which does not calculate the whole graph. It only considers the influence of surrounding nodes on the node, and the calculation is faster.

Step 5: Model prediction and accuracy evaluation

Based on the trained model in Step 4, we perform functional prediction for unlabeled streets in study area. Since we introduce a semi-supervised machine learning approach, we include only a fraction of the labels of the data. We borrow the accuracy criterion from recommendation systems and measure the model accuracy by the $P@K$ score (Ge et al., 2010), which calculates the proportion of the nearest K regions that correctly contain the predicted labels. As such, $P@K$ denotes when the predicted street label in the test data appears in the $nearest - k$ true region label list (the EULUC-China dataset in Step 3 is regarded as the actual value). Num denotes the number of test data, and n_{oc} denotes the number of times this occurs, as follows:

$$P@K = \frac{n_{oc}}{Num} \tag{6}$$

A smaller K value means a stricter evaluation criterion, and a more extensive P value indicates a higher model accuracy and a closer distribution to the actual condition. Since the study area contains more than 5,000 streets and 2,000 regions, we use KD-Tree to speed up the retrieval and calculation.

3. Experiment

3.1. Introduction to the study area and data pre-processing

As shown in Figure 5, this research is divided into three main parts. The first part is the cross-modal decoder of the streetscape images to obtain a highly semantic compressed description of the city scene; the second part is the semantic encoder of the decoder results together with the OSM network topological relations and then input to the graph network for semi-supervised learning; finally, we extract the knowledge triple from the images to generate a street semantic knowledge graph and text summaries under each city function, which make the city not only can be 'watched' but also can be 'read'.

Our study area is within the third ring of Wuhan in China, we collected 75,628 Tencent Street View images, which

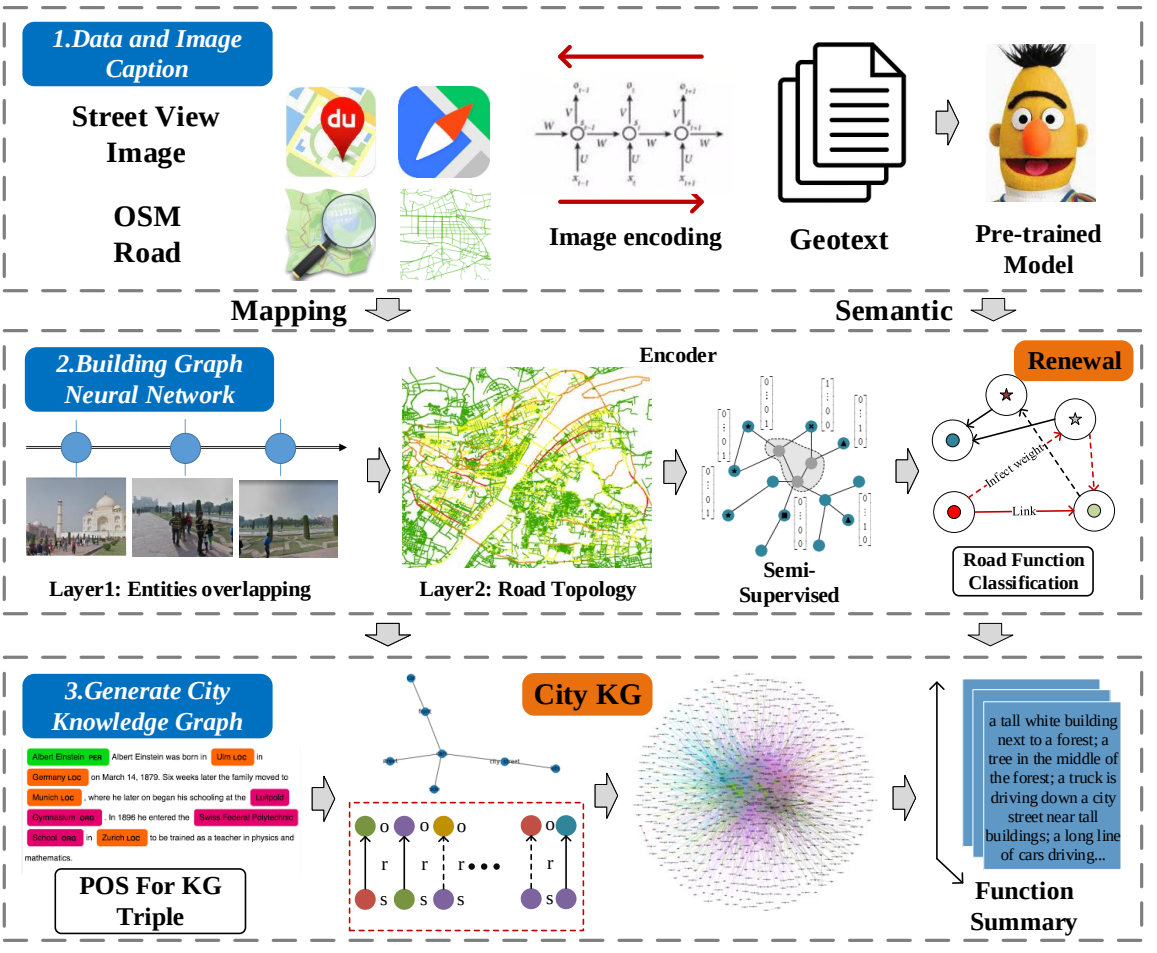


Figure 5: General structure of the approach (Step 1 – Data process and Image Caption; Step 2 – Build graph neural network and train it; Step 3 – Extract triple and Generate Knowledge Graph). Source of the map: © OSM contributors.

were sampled at a similar time to the launch of the EULUC-China product, while 64,750 newly (collected in the July of 2022) Baidu Street View images were used to sense the recent urban construction in the study area. By calculating images from different times, we explore the urban functional transfer and the validity of the method from different data sources.

There are many diverse urban scenes and physical entities in the study area, including 5,458 OSM roads and 18,907 streetscape sampling locations. Based on the pre-trained visual language model mentioned in Figure 4 Step 1, we perform cross-modal decoding of 75,628 street scene images in the study area and generate the most relevant five scene descriptions for each. Consequently, we obtained a total of 378,140 sentences after about 72 hours computation on a Tesla P100 GPU.

3.2. Model training and accuracy comparison

We calculated the 25-meter buffer of OSM roads (street scenes are collected by sampling vehicles, and most urban roads are no more than 50 metres wide), counted the streetscapes that fall within the buffer zone as a representation of the street, built the 28,693 spatial connection between OSM and Sreet View. And then we use the EULUC to label the streets. We continued the following research based on these information.

We use Queen contiguity to calculate the spatial weight matrix, with 23,549 neighboring relationships. There are 5,458 OSM segments in the study area, with an average of 5 relations per street, so we calculate 5-Nearest (K=5) neighborhoods (27,290 neighboring relationships) as another set of spatial weights.

Table 3
The accuracy comparison of 12 models under different conditions.

Model	Encoder	Weight	Graph Sructure	@1	@3	@5	@10	Number	Time(s)
Model 1	Bert	K-Nearest	GCN1	0.344	0.568	0.654	0.758	8	39.672
Model 2	Bert	K-Nearest	GCN2	**0.395**	**0.638**	0.73	0.807	10	27.899
Model 3	Doc	K-Nearest	GCN1	0.327	0.544	0.632	0.738	9	24.516
Model 4	Doc	K-Nearest	GCN2	0.327	0.548	0.64	0.75	10	18.402
Model 5	Bert	Queen Contiguity	GCN1	0.382	0.637	**0.736**	**0.824**	7	32.374
Model 6	Bert	Queen Contiguity	GCN2	0.321	0.554	0.648	0.752	10	21.209
Model 7	Doc	Queen Contiguity	GCN1	0.335	0.564	0.657	0.75	8	24.298
Model 8	Doc	Queen Contiguity	GCN2	0.289	0.496	0.593	0.698	10	17.579
Model 9	Bert	K-Nearest	GAT	**0.462**	**0.730**	**0.824**	**0.899**	10	218.722
Model 10	Bert	Queen Contiguity	GAT	0.376	0.623	0.723	0.818	10	204.958
Model 11	Bert	K-Nearest	GraphSAGE	0.364	0.592	0.685	0.774	10	36.012
Model 12	Bert	Queen Contiguity	GraphSAGE	0.376	0.623	0.723	0.818	10	32.539

Note: GCN1,GCN2: the computational units is 64 and 32, 32 and 16 respectively

We conducted 12 sets of comparison experiments using different models as shown in Table 3. The main differences between the models are the different semantic encoder, the spatial weight matrix and the graph feature extraction network, and each model is given a unique name (Model1, Model2, etc.). We performed 200 rounds of epochs for all 12 models to test their prediction accuracy, and the training loss variations are shown in Figure 6. We only discuss models where the number of prediction categories is equal to 10, because these models can maintain a better generalization ability. As can be observed, using Bert as the encoder will achieve higher accuracy (Model 2 and Model 4). Such high accuracy shows that Bert has a solid ability to extract scene features and can reduce training loss. On the contrary, the traditional method (Doc2vec) can not reduce the loss even after 200 epochs. Take model 2 for example, when using the basic GCN structure as a semi-supervised classifier, P@1=0.395, P@3=0.638, P@5=0.730, and P@10=0.807. The result means that under the stricter criterion (P@1), about 40% of the predicted labels of streets agree with the EULUC; under the loose criterion (P@10), about 80% of the predicted street labels are consistent with one or more labels of the ten nearest parcels. Note that this does not mean that the accuracy of our model is only 40%. Limited by the low quality of EULUC data and since one street may have multiple functions, loose criterion (P@10) may be a better choice of model assessment.

Mentioned in Step 2 of the method, GAT requires more parameters for training (additional weight matrix), Model 9 and Model 10 are the slowest but achieve the highest accuracy (P@1=0.462, P@3=0.730 P@5=0.824, P@10=0.899). This performance is quite satisfactory and the model's predictions are excellent. The performance proves that Street View can detect urban street functions effectively, and introducing spatial topology improves the city's interpretation level.

In addition, as shown in Figure 6 that when GCN is used as a classifier (Model 2, Model 3), the training loss is low at the beginning, indicating that the model already has a good feature extraction capability without training. Model 2, Model 6, Model 9, Model 11, and Model 12 have the fastest converge speed. These models also have good generalization ability and can accurately identify the ten functional categories. Using the 5-Nearest (K=5) neighborhoods (dashed line, Model 5, Model 6) to calculate the spatial weights makes the model more likely to converge than Queen contiguity (solid line, Model 1, Model 2). The results in Table 3 also show the lower accuracy of using Queen contiguity.

We use the T-distributed stochastic neighbor embedding (T-SNE) dimensionality reduction method to visually check if the model can learn the classification features well. As we can see in Figure 7, the 768-dimensional features of the street nodes are reduced to two dimensions, and the labels of the model predictions are drawn. Figure 7 can visually verify the algorithm's effectiveness to see whether the model can learn similar embeddings of nodes belonging to the same category. More specific, we can use T-SNE method for processing the feature embedding ($length = 10$, number of land use labels), which is the last layer output of the neural network. At this time, the embedding has been calculated by forward propagation of the trained neural network. Besides, the predicted labels can also be obtained after the calculation of the maximum value function (argmax).

In addition to the within-group comparisons performed in Table 3, we also compared the results by the traditional

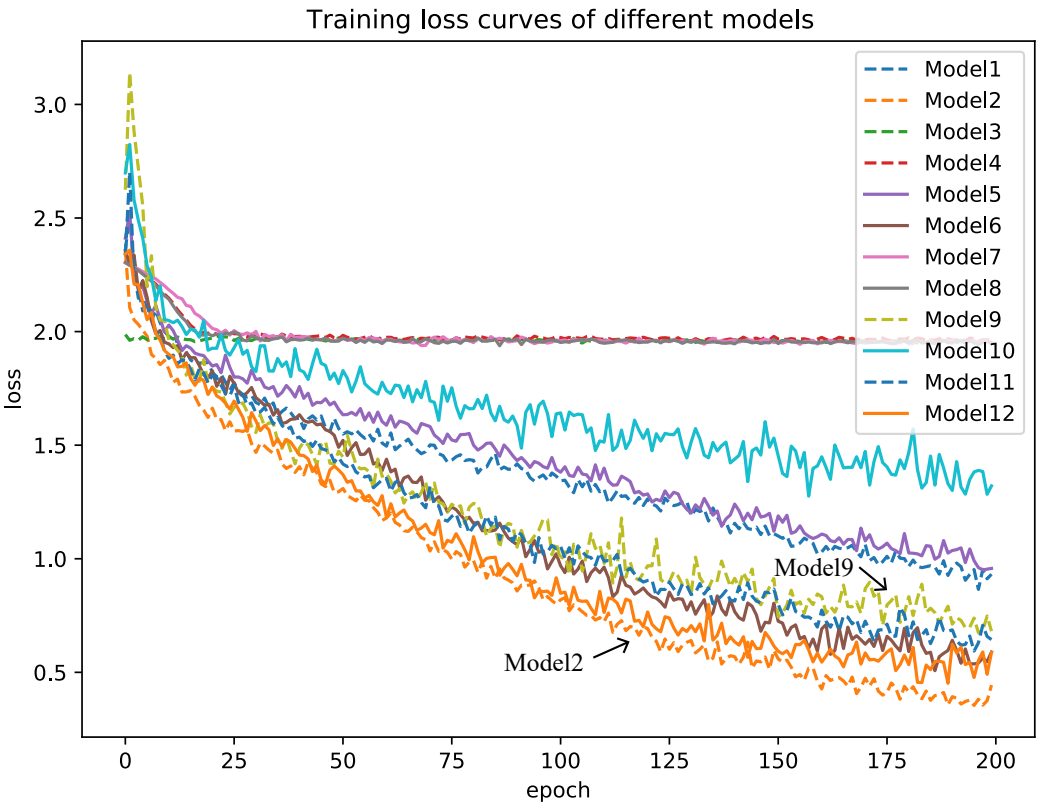


Figure 6: Training loss curves of different models. Dashed lines indicate the use of 5-Nearest neighborhoods, and solid lines indicate the use of Queen contiguity.

machine learning method Multilayer Perception (MLP), without considering the spatial topology relationship. As shown in Figure 7, Figure 7 (a) shows the predicted feature vector dimensionality reduction results for model 9, Figure 7 (b),(c) show the MLP feature vector dimensionality reduction results when the neurons are 32 and 128, respectively, and the Bert is also used as the semantic encoder. Figure 7 (d) shows the case when a two-layer neural network is used. we can see the tradition method is much less effective than our method, with some classes missing from the predictions (only four categories). The advantage of our approach over traditional methods is that our forecasts take into account not only the current street environment but also the environment of the adjacent streets. In addition, we also use the test set to independently evaluate the performance of the model by calculating the number of correctly predicted nodes (roads or streets) as the percentage of the test set number. The models represented by Figure 7(a,b,c,d) achieved accuracies of 0.844, 0.445, 0.624 and 0.695, respectively, and our method remains optimal after inter-group and intra-group comparisons.

The labels in Figure 7 are based on the categories predicted by the model, and the purpose of the figure is to visualize how different models perform in separating different categories. Poorly generalized models do not accurately identify all categories (predicted category number less than 10). One reason is that some categories have less training data, which makes it difficult for non-graphical models to fully learn the features of each category. This is particularly challenging for models that rely on large amounts of data, as there is not enough information for them to learn. Lack of data can lead to poor generalization of the model, which results in reduced accuracy for some categories.

This phenomenon can be observed in Figure 7, where some models fail to predict all categories correctly and certain categories are less accurate than others. The same can be seen in the table 3.

It is worth noting that this is a common challenge in machine learning, especially when dealing with unbalanced datasets. The method proposed in this paper introduces graph neural networks to improve this semi-supervised (not rich in training data) situation, which are able to process continuous street images as input and fully take into account the spatial dependencies between road networks, which makes it more robust when dealing with less data and better

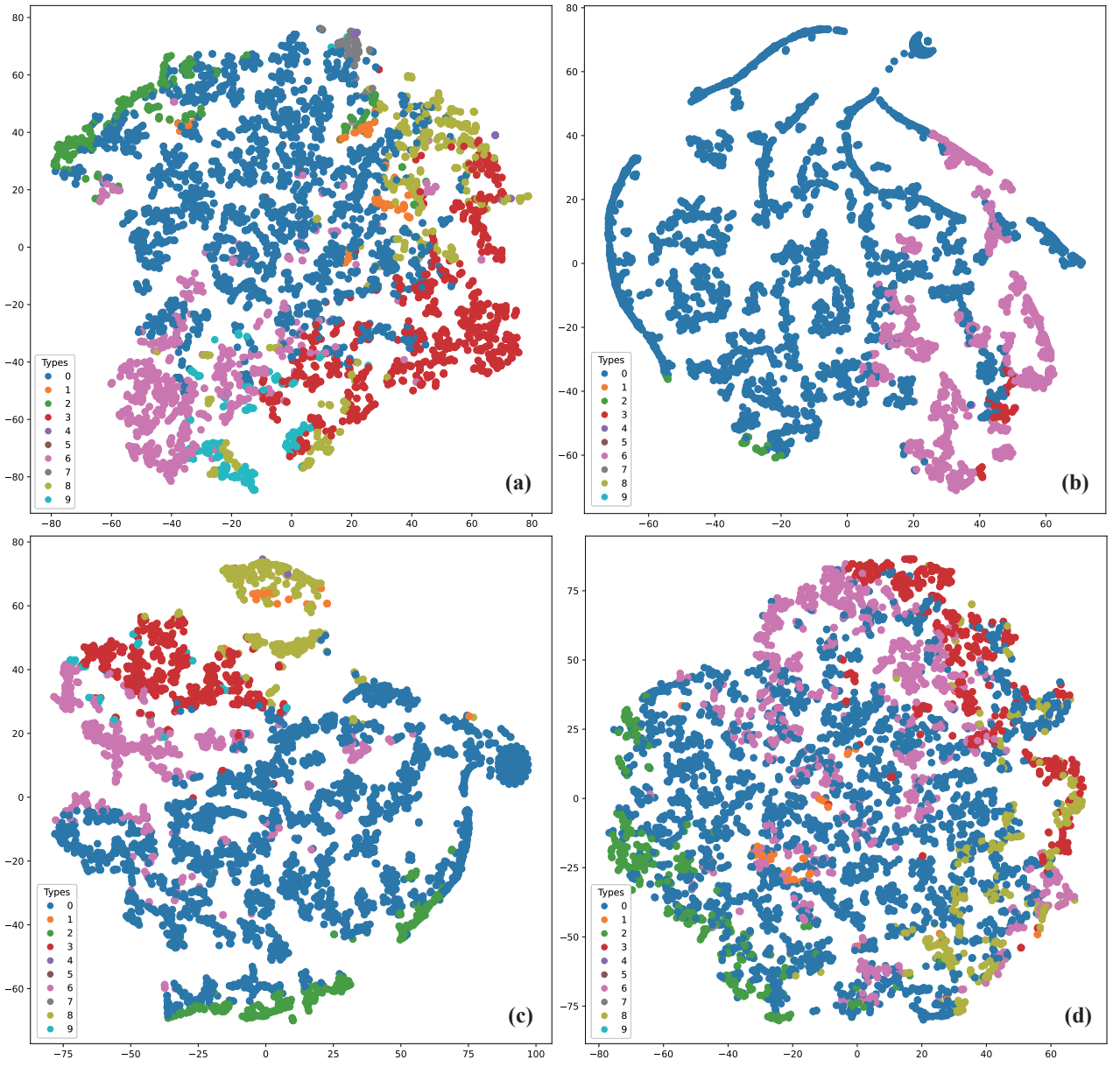


Figure 7: T-SNE dimensionality reduction results of our method (a) compared to traditional methods (b,c,d).

generalizes to unseen data.

In addition, the different classes in Figure 7(b) show uneven spatial distributions due to the nature of the T-SNE algorithm, which tries to preserve pairwise distances between data points in high-dimensional space, but in low-dimensional space, distances between points are not guaranteed to be preserved. In addition, the uneven distribution may also be due to an imbalance in the data, i.e., some classes are more frequent than others, which may cause the algorithm to focus more on these classes and create a denser cluster (Figure 7 c), while other classes may be underrepresented in the graph.

The confusion matrix of Model 9 on the training data is shown in Table 4. Model 9 has an excellent fitting ability to identify all ten urban function classes. However, almost every category has a portion of streets misclassified as residential since the residential class is the most common type in the city and is often mixed with other functions.

Table 4
The training confusion matrix of Model 9.

Types	Res	Bus	Com	Ind	Tra	Adm	Edu	Med	Spo	Par	PredNum
Residential	85	0	1	0	0	0	0	1	0	0	3392
Business	5	9	1	0	0	0	0	0	0	0	158
Commercial	8	1	20	0	0	0	0	0	0	0	293
Industrial	13	0	1	21	0	0	1	0	1	0	412
Transportation	1	0	0	0	5	0	0	0	0	0	34
Administrative	2	0	0	0	0	4	0	0	0	0	48
Education	13	0	0	0	0	0	52	0	0	0	747
Medical	1	0	0	0	0	0	0	11	0	0	37
Sports/Culture	4	0	1	0	0	0	0	0	23	0	248
Parks	2	0	0	0	0	0	0	0	0	8	89

Therefore, the residential category has the highest percentage in 3,392 streets in the prediction results of Model 9 (Table 4). In contrast, the education, transportation, and administrative categories have the lowest number, with less than 50 streets identified.

3.3. Model prediction and node attribute discussion

The Model 9 prediction results with the highest accuracy are shown in Figure 8. It can be seen that the prediction results are generated at a fine scale, and the labels 0 to 9 are Residential, Business Office, Commercial Service, Industrial, Transportation Station, Administration, Education, Medical, Sports and Culture, Parks and Green Space. As shown in the Figure 8 subgraph, streets are the "nerve endings" of the city. Our model can identify actual and different functions despite the short distance between two streets. Compared with using building footprint or TAZ as the spatial unit, we can locate more rich results and are also good at exploring multiple functions of the same region.

Figure 9 (a-j) shows the 10 scenes predicted as categories 0-9 respectively. Figure 9 (a) is identified as Type 0 (Residential), with the most noticeable feature being tall residential buildings. Figure 9 (b) is identified as Type 1 (Business Office), the most apparent feature is the office building and a large number of parked cars. Figure 9 (c) is identified as Type 2 (Commercial Service). As indicated in the figure, the street is busy, and there are superstores with red signs on both sides. Figure 9 (d) is identified as Type 3 (Industrial), white vans can be seen parked on the street in front of a factory. Figure 9 (e) is identified as Type 4 (Transportation Station). It can be seen from the figure that a bus parking lot parked with many public buses. Figure 9 (f) is identified as Type 5 (Administrative), and the identification results seem to be highly correlated with Overpass. The condition may be because the model does not learn the features of this class sufficiently (only six streets are labeled in Table 2), and there may be errors in the labeled data. On the other hand, Type 5 street may be challenging to distinguish effectively from other classes (e.g. Residential) in terms of visual features. Figure 9 (g) is identified as Type 6 (Education). The main difference between it and Figure 9 (j, Type 9, Parks and Green Space) is that the Type 6 functional street also includes buildings (dormitories, academic buildings) with vehicles and pedestrians, while Figure 9 (j) includes only the natural environment. Figure 9 (h) is identified as Type 7 (Medical). It is difficult to observe whether it is correctly predicted at this location using those four different angle images, which is also related to the fact that the Type 7 also contains fewer visual features. Figure 9 (i) is identified as Type 8 (Sports and Culture), where we can see the presence of basketball courts on both sides of the road, and the model keenly identifies this phenomenon.

In addition, since the model has a solid ability to identify urban open spaces (Figure 9 (a-j)), it shows that the street view can not only extract information about the streets but also can detect the functions of the regions on the roadsides. In addition, this study demonstrates that streetscape also has the potential to be used as an independent data source to identify urban functions.

We mentioned that Figure 9(f) illustrates the model misclassifies the types of urban functions without distinctive visual features. This phenomenon occurs because there are fewer differences in visual elements between medical, administrative, and residential. We calculated the mean of the embedding vector $Embd\_Type_i$ (Equation 7) for each city function as an overall expression and calculated the cosine similarity between different function categories (Figure 10).

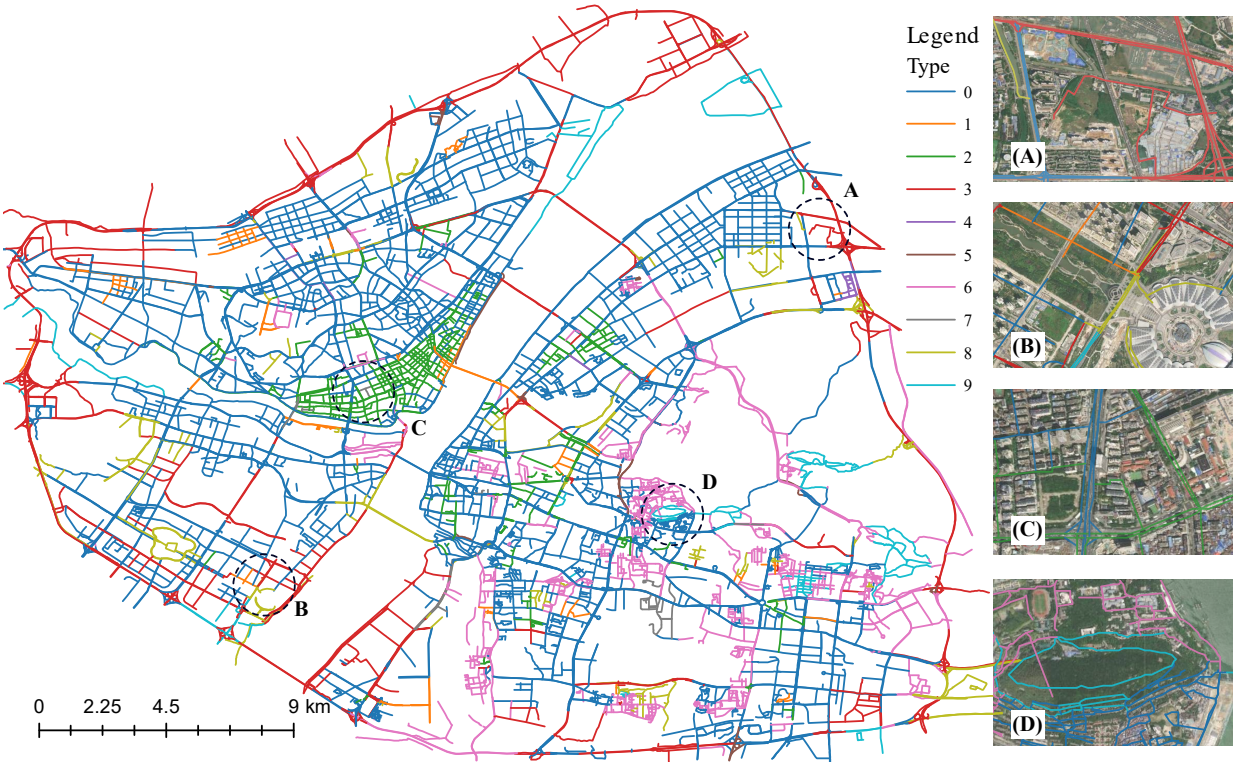


Figure 8: Prediction results of SOTA model in the study area. Source of the road data: OpenStreetMap contributors. Source of the satellite imagery: USGS/NASA Landsat.

Figure 9: Randomly selected 10 locations predicted by the model 9 for categories 0-9 (each street has dozens of images, we randomly selected 4 angles of a location for illustration). Source of the imagery: Tencent Street View.

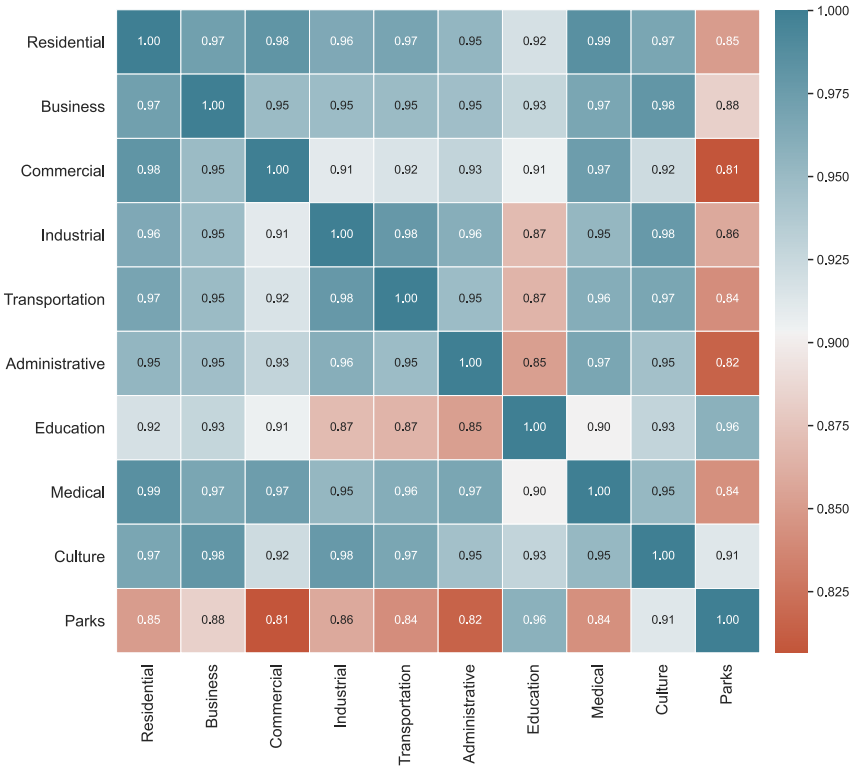


Figure 10: The cosine similarity between different function categories based on Equation 7. The semantic mean embedding of each class is used as the class overall representation.

$$Embd\_Type_i = \frac{\sum_{j=1}^{Num_i} Embd\_Street_j}{Num_i} \tag{7}$$

The semantic similarity between the different urban functional classes is high, with most similarities above 0.8. It is difficult for traditional machine learning models to distinguish the different categories without incorporating the dual spatial dependencies of knowledge and topology (Figure 7). The differences between Education, Industrial, and Transportation categories are great. Park and Green Space, on the other hand, is more similar to Education and Sports and Culture because these areas are often greener and share some characteristics (e.g., beautiful surroundings, less traffic, and pedestrians). Residential differs more from the Education and Park. As residential areas are common functional areas in cities, they have some spatial co-occurrence with most categories, with the slightest difference with the medical category (correlation coefficient = 0.99). The Commercial and Park categories, one being an area of heavy traffic and economic prosperity and the other being a more sparsely populated area, have the most significant difference, with a correlation coefficient of 0.81.

3.4. Urban renewal and Spatial-Temporal changes

We collected the Baidu Street View for the same study area in July 2022, which reflects newer urban conditions. Tencent Street View (which we used in the previous section), on the other hand, reflects the older urban conditions and has been updated less frequently since 2014.

To capture more real urban spatial-temporal changes, we set Baidu Street View's sampling location and angle consistent with Tencent Street View. Although Baidu Maps do not cover some places, we collected 64,750 Baidu street images. We adopted the same processing flow as shown in Figure 3, using the parameter settings of the SOTA model

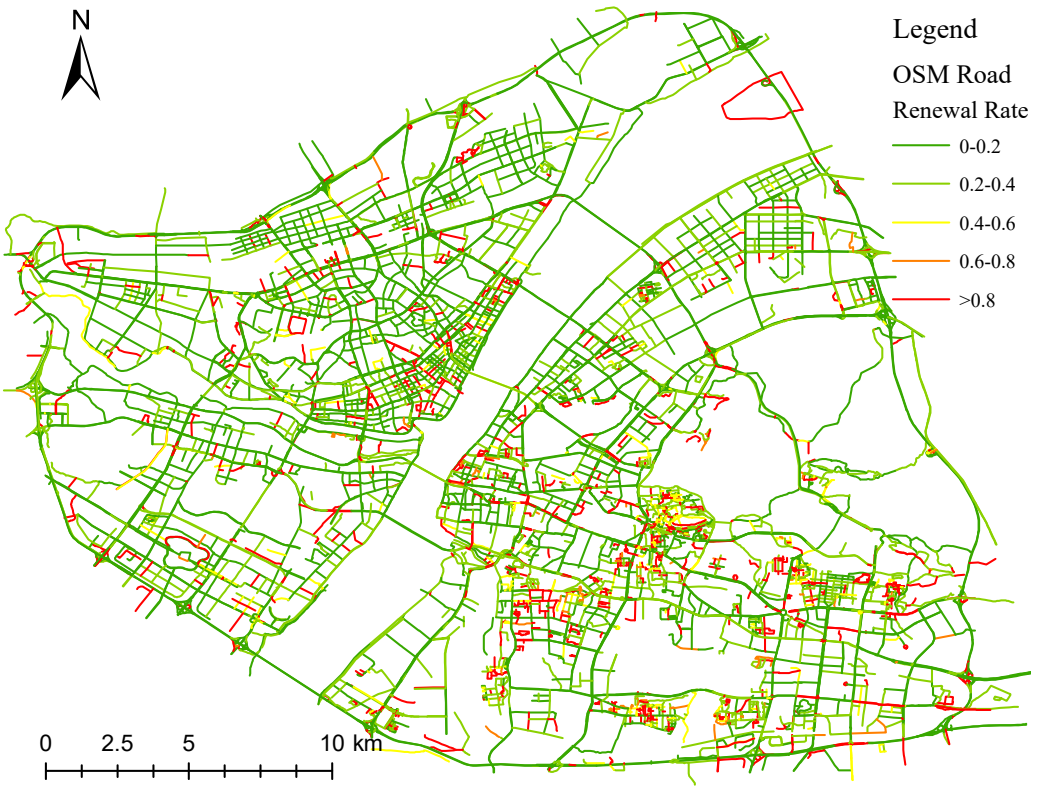


Figure 11: Urban renewal rate distribution based on Equation 8. Source of the road network: © OSM contributors.

(Model 9). We obtained the embedded representation of each street (Step 2) and the new urban function distribution (Step 5).

Based on the street embedding generated in Step 2 with different data, we calculate the street renewal rate based on Equation 8, which is used to analyze the urban function evolution trend in recent years.

$$Renewal\_Rate = 1 - Cosine\_Similarity(Embedding_{Tencent}, Embedding_{Baidu}) \tag{8}$$

In Figure 11, most of the areas, especially the main roads in the city, have changed less. Meanwhile, some areas have changed significantly (renewal rate>0.8). Furthermore, we plot the overall urban function transfer matrix in Table 5.

The transfer matrix is a two-dimensional matrix calculated based on the relationship between the changes in urban functions in different phases of the same area. It includes the inter-conversion between different land types and the overall change trend. Among all the 5,458 streets, the residential type has the largest decrease (1877 streets reduced and 433 new streets added), reflecting the gradual demise of urban villages in recent years supported by the china 'shantytown renovation' policy. The net increase in commercial and industrial streets reflects the increasingly dynamic economy and industry development. Cultural and green spaces show a non-significant shrinkage. Wuhan is the city with the largest number of lakes in China. The continued anthropogenic activities and urbanization have led to serious degradation of ecological wetlands, which is also consistent with other's studies (Wang et al., 2020).

3.5. Street Semantic Knowledge Graph Construction

The method Step 1 mentions that we generate scene descriptions based on entity relations (s->r->o). We extract knowledge triples from scene descriptions based on Spacy (Lai et al., 2022). This natural language processing tool library can identify the part-of-speech of words and the dependencies between words (Fang et al., 2021). For simplicity, we extract entities from words with the lexicality 'NOUN' and relations only from words with 'VERB' and 'ADP'. For

Table 5
The city function transfer matrix of Model 9.

Old \ New	Res	Bus	Com	Ind	Tra	Adm	Edu	Med	Spo	Par	Sum	Decrease
Residential	1515	12	591	636	11	15	460	12	124	16	3392	1877
Business	66	22	12	27	5	1	21	0	1	3	158	136
Commercial	91	5	160	4	0	0	31	2	0	0	293	133
Industrial	59	0	3	278	2	0	49	0	19	2	412	134
Transportation	4	0	2	12	15	0	1	0	0	0	34	19
Administrative	6	0	8	15	4	13	2	0	0	0	48	35
Education	154	0	18	76	0	0	486	0	5	8	747	261
Medical	7	0	10	2	0	0	0	18	0	0	37	19
Sports/Culture	36	0	27	111	0	2	19	0	52	1	248	196
Park Green Space	10	1	2	16	0	0	30	0	1	29	89	60
Sum	1948	40	833	1177	37	31	1099	32	202	59	5458	
Add	433	18	673	899	22	18	613	14	150	30		2870

example, we can extract the triples of, s->van,r->parked,o->side, s->van,r->parked,o->street, etc from the sentence of "a white van parked on the side of a street".

We perform similar calculations for all scenes and obtain 75,628 urban knowledge triples. Based on these knowledge triples, we generate the semantic knowledge graph of all streets in Wuhan as shown in Figure 12. It includes 707 urban environment entities and 4,371 spatial location relationships. The larger nodes indicate the higher frequency of the entity, and the thickness of the edges indicates the strength of the correlation relationship. After that, our graph-based community detection method (Louvain method) (Traag et al., 2019; Zhang et al., 2022b) divides these entities and relationships into multiple communities, and entities co-occur more frequently with entities inside the community than with entities outside the community, which we do not explain in-depth here.

Based on the street labels predicted by Model 9, we perform similar processing for each class of urban functions to identify the differences between them. Since we have predicted ten categories, we choose only three here as examples, Commercial, Education, Park and green space, containing 259 nodes with 710 edges, 400 nodes with 1,257 edges, and 209 nodes with 625 edges, respectively (Figure 13, namely, city function knowledge graph). More nodes indicate richer urban entities within this function area, and more edges indicate stronger spatial relationships between entities.

Buildings, stores, road, etc. appear frequently in Figure 13(a), while tree appears more frequently in Figure 13(b,c). The frequency of sidewalk and forest entities in Figure 13(c) varies more considerably from other categories. We can get above information from these sub-graphs clearly, demonstrating the effectiveness of the street view semantic-based approach in identifying urban functions.

In addition, we compute 'summaries' of each function description set based on the method provided by HuggingFace open source community (Wolf et al., 2020) to generate a 'sentence' that best represents it. These sentences consist of the most important and most characteristic scene descriptions. Function summary is shown below:

- **Commercial:** a street with a **large building** in the background; a red car driving down a street next to **tall buildings**; a view of a city street filled with tall buildings in the distance; **a group of cars** that are sitting in the street; a view of a city street from a distance. a city road filled with **traffic surrounded** by tall buildings; a highway filled with lots of traffic and tall buildings.

- **Education:** a tall white building next to a **forest**; a tree in the middle of the **forest**; a truck is driving down a city street near tall buildings; a long line of cars driving down a city street; a sign that is on the side of a building; a large building with a **building** in the background; a large blue **bench** sitting in front of a city skyline. A group of cars that are sitting in the street; a boat that is sitting in the dirt ;

- **Park and Green Space:** an **empty street** with **no cars** on it; a highway with cars driving down the road; a blue car parked in a parking lot. a view of a bridge from across the water; an **empty** highway with a sign; a long road with a long line of traffic lights on it; a car is driving down a road next to a bridge; a train is traveling over a bridge over a **river**; a small **tree** on the side of the road; a group of cars that are sitting in the street.

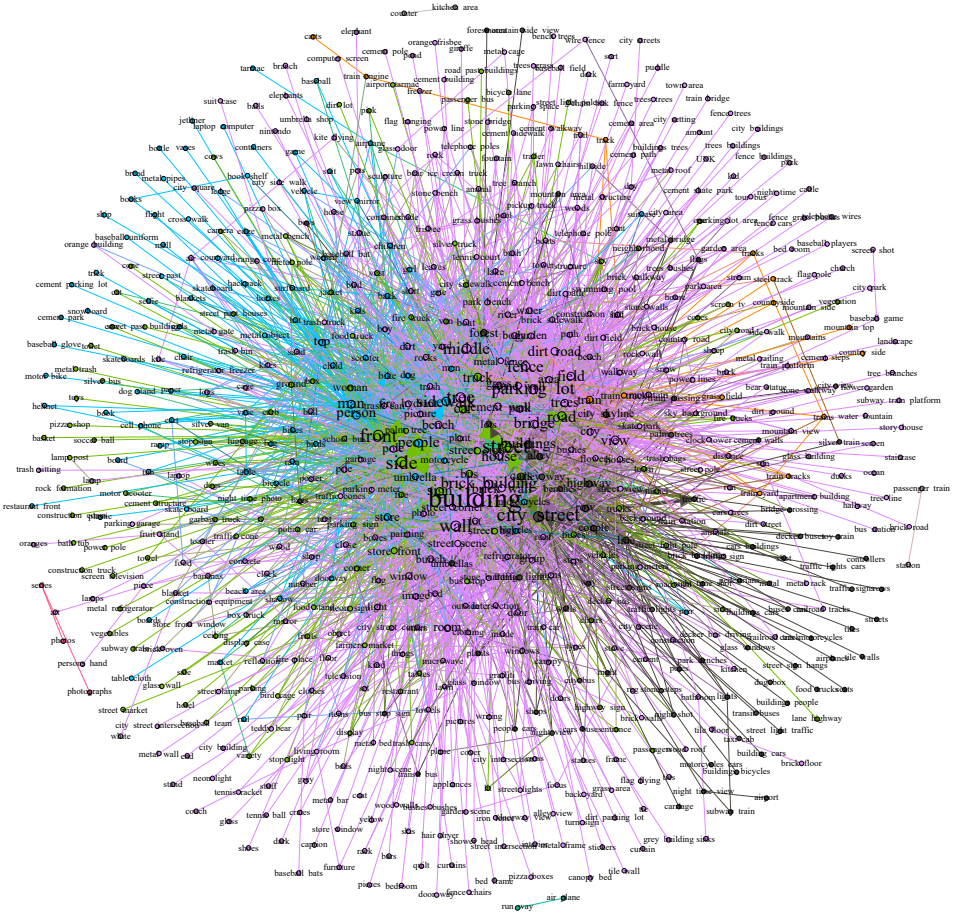


Figure 12: Wuhan Street Semantic Knowledge Graph (nodes represent the environmental entities of the city, edges represent the spatial relationships between entities, and the colors represent different 'communities').

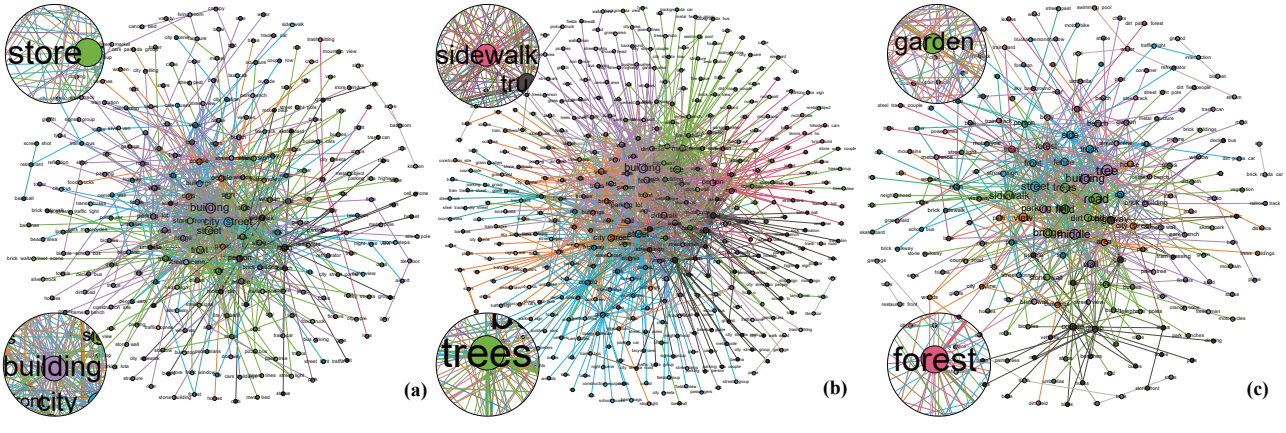


Figure 13: Comparison of city function knowledge graph (a,b,c represent Commercial, Education, Park and green space respectively).

As we can observe, the results of the city function 'summary' are similar to the city function knowledge graph, but they are easier to read, and the characteristic elements can be found more easily.

4. Conclusion and discussion

Urban geo-tagged proximate sensing images are generated continuously in cities. For example, they include street view imagery and photos shared on social media. They may be considered as twin mappings of the city's operational state, constantly 'refreshing' for sensing urban areas. Faced with a large volume of geo-tagged images, capturing the rich semantic information and the spatio-temporal location relationship is crucial to understanding and interpreting urban space. In this paper, we proposed a purely visual scheme for the functional perception of urban streets, which incorporates urban knowledge and road network topology and can fuse multiple source images to generate a holistic representation of a spatial unit. We also incorporate temporal information and integrate historical street images to calculate urban spatial-temporal changes, renewal rates, and urban function transition matrix.

Our method has three significant novelties and advantages: fusing multiple sources of urban proximate sensing images, supporting arbitrary scale spatial units, and being rich in socio-economic information. The first advantage is that we can include the street view and other proximate images in the computation and generate the semantic representation (city caption), as long as the pictures can show the actual conditions of the places, and even indoor images can be included. These images can collectively form the overall image of the city (Filomena et al., 2019), driving our understanding of the space deeper. The second advantage is that our research unit is not only limited to the street scale but also generalisable to other scales, even for a small sampling point or a region. The different scales differ only in the length of the semantic description text and the spatial topological relations. The third advantage is that because the city is rich in human activity footprints, the proximate images can penetrate deep into the city and provide real-time feedback on the 'people', 'vehicles', and 'things', and the raw shape & color information. It will help to sense the socio-economic environment comprehensively. In contrast, remote sensing images in urban space are difficult to have a such high spatial and temporal resolution.

This approach still has some limitations and elements that need to be discussed. The primary drawback is the low scalability (3 days for the Step 1 computations) because the cross-modal step (generating caption) is computationally-intensive. In addition, the maximum length of BERT is limited to 512 words. The generated spatial unit description text can be long if the selected spatial unit is too large and with excessive images. The model cannot get all the information thoroughly. One can choose the text truncation or based on the introduced text summarization method before processing. It is important to note that this new geo-intelligent analysis method can be used not only for urban function recognition but also for capturing and classifying socio-economically rich and sensitive features, such as urban vibrancy, urban village identification, built-up area recognition, etc. In addition, this paper mainly discusses the street view images provided by commercial and crowd-sourced services, which have more uniform spatial distribution and longer update intervals. Understanding how to use the social sensing (user-generated) (Liu et al., 2015) images which tend to be unevenly spatially distributed and got posted much more rapidly. Testing the applicability of our model will be the focus of our following research.

Acknowledgements

We thank our colleagues at the NUS Urban Analytics Lab for the discussions and thank the editor and reviewers for their professional comments. This research is supported by (i) the National Key R&D Program (no.2018YFB2100500), (ii) the National Nature Science Foundation of China (no. 41971351), and (iii) the Singapore Ministry of Education Academic Research Fund Tier 1 (project Multi-scale Digital Twins for the Urban Environment: From Heartbeats to Cities).

References

Abdelrahman, M.M., Miller, C., 2022. Targeting occupant feedback using digital twins: Adaptive spatial–temporal thermal preference sampling to optimize personal comfort models. Building and Environment 218, 109090. doi:10.1016/j.buildenv.2022.109090.

Amiruzzaman, M., Curtis, A., Zhao, Y., Jamonnak, S., Ye, X., 2021. Classifying crime places by neighborhood visual appearance and police geonarratives: a machine learning approach. Journal of computational social science 4, 813–837.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6077–6086.

Biljecki, F., Chow, Y.S., 2022. Global Building Morphology Indicators. Computers, Environment and Urban Systems 95, 101809.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. Landscape and Urban Planning 215, 104217.

Bruna, J., Zaremba, W., Szlam, A., LeCun, Y., 2014. Spectral networks and locally connected networks on graphs, in: Bengio, Y., LeCun, Y. (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. URL: http://arxiv.org/abs/1312.6203.

Campbell, A., Both, A., Sun, Q.C., 2019. Detecting and mapping traffic signs from google street view images using deep learning and gis. Computers, Environment and Urban Systems 77, 101350.

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 Journal of Photogrammetry and Remote Sensing 163, 82–97.

Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., Zhang, Q., Qiu, G., 2018. Integrating aerial and street view images for urban land use classification. Remote Sensing 10, 1553.

Chadzynski, A., Li, S., Grisiute, A., Farazi, F., Lindberg, C., Mosbach, S., Herthogs, P., Kraft, M., 2022. Semantic 3d city agents—an intelligent automation for dynamic geospatial knowledge graphs. Energy and AI 8, 100137.

Chen, B., Tu, Y., Song, Y., Theobald, D.M., Zhang, T., Ren, Z., Li, X., Yang, J., Wang, J., Wang, X., et al., 2021a. Mapping essential urban land use categories with open big data: Results for five metropolitan areas in the united states of america. ISPRS Journal of Photogrammetry and Remote Sensing 178, 203–218.

Chen, L., Lu, Y., Sheng, Q., Ye, Y., Wang, R., Liu, Y., 2020. Estimating pedestrian volume using street view images: A large-scale validation test. Computers, Environment and Urban Systems 81, 101481.

Chen, L., Lu, Y., Ye, Y., Xiao, Y., Yang, L., 2022. Examining the association between the built environment and pedestrian volume using street view images. Cities , 103734.

Chen, W., Wu, A.N., Biljecki, F., 2021b. Classification of Urban Morphology with Deep Learning: Application on Urban Vitality. Computers, Environment and Urban Systems 90, 101706.

Chen, Y., Liu, X., Li, X., Liu, X., Yao, Y., Hu, G., Xu, X., Pei, F., 2017. Delineating urban functional areas with building-level social media data: A dynamic time warping (dtw) distance based k-medoids method. Landscape and Urban Planning 160, 48–60.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., Karagiorgou, S., Efentakis, A., Lamprianidis, G., 2015. Crowdsourcing urban form and function. International Journal of Geographical Information Science 29, 720–741.

Fang, F., Yu, Y., Li, S., Zuo, Z., Liu, Y., Wan, B., Luo, Z., 2021. Synthesizing location semantics from street view images to improve urban land-use classification. International Journal of Geographical Information Science 35, 1802–1825.

Filomena, G., Verstegen, J.A., Manley, E., 2019. A computational approach to 'the image of the city'. Cities 89, 14–25.

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

Ge, M., Delgado-Battenfeld, C., Jannach, D., 2010. Beyond accuracy: Evaluating recommender systems by coverage and serendipity, in: Proceedings of the Fourth ACM Conference on Recommender Systems, Association for Computing Machinery, New York, NY, USA. p. 257–260. URL: https://doi.org/10.1145/1864708.1864761, doi:10.1145/1864708.1864761.

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., Kalogirou, S., 2021. Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto International 36, 121–136.

Gong, P., Chen, B., Li, X., Liu, H., Wang, J., Bai, Y., Chen, J., Chen, X., Fang, L., Feng, S., et al., 2020. Mapping essential urban land use categories in china (euluc-china): Preliminary results for 2018. Science Bulletin 65, 182–187.

Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. Advances in neural information processing systems 30.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Hong, D., Yokoya, N., Xia, G.S., Chanussot, J., Zhu, X.X., 2020. X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data. ISPRS Journal of Photogrammetry and Remote Sensing 167, 12–23.

Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H., 2019. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) 51, 1–36.

Hu, C.B., Zhang, F., Gong, F.Y., Ratti, C., Li, X., 2020a. Classification and mapping of urban canyon geometry using google street view images and deep multitask learning. Building and Environment 167, 106424.

Hu, S., Gao, S., Wu, L., Xu, Y., Zhang, Z., Cui, H., Gong, X., 2021. Urban function classification at road segment level using taxi trajectory data: A graph convolutional neural network approach. Computers, Environment and Urban Systems 87, 101619.

Hu, S., He, Z., Wu, L., Yin, L., Xu, Y., Cui, H., 2020b. A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. Computers, Environment and Urban Systems 80, 101442.

Huang, X., Yang, J., Li, J., Wen, D., 2021. Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery. ISPRS Journal of Photogrammetry and Remote Sensing 175, 403–415.

Inoue, T., Manabe, R., Murayama, A., Koizumi, H., 2022. Landscape value in urban neighborhoods: A pilot analysis using street-level images. Landscape and Urban Planning 221, 104357.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS journal of photogrammetry and remote sensing 145, 44–59.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations (ICLR).

Lai, K., Porter, J.R., Amodeo, M., Miller, D., Marston, M., Armal, S., 2022. A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures. Information Processing & Management 59, 102735.

Lauko, I.G., Honts, A., Beihoff, J., Rupprecht, S., 2020. Local color and morphological image feature based vegetation identification and its

application to human environment street view vegetation mapping, or how green is our county? Geo-spatial Information Science 23, 222–236.
Li, X., Hu, T., Gong, P., Du, S., Chen, B., Li, X., Dai, Q., 2021a. Mapping essential urban land use categories in beijing with a fast area of interest (aoi)-based method. Remote Sensing 13, 477.
Li, Y., Kong, D., Zhang, Y., Tan, Y., Chen, L., 2021b. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. ISPRS Journal of Photogrammetry and Remote Sensing 179, 145–158.
Li, Y., Zhang, X., Gu, J., Li, C., Wang, X., Tang, X., Jiao, L., 2021c. Recurrent attention and semantic gate for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing 60, 1–16.
Lin, L., Wang, G., Zuo, W., Feng, X., Zhang, L., 2016. Cross-domain visual matching via generalized similarity measure and feature learning. IEEE transactions on pattern analysis and machine intelligence 39, 1089–1102.
Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.
Liu, B., Deng, Y., Li, M., Yang, J., Liu, T., 2021. Classification schemes and identification methods for urban functional zone: A review of recent papers. Applied Sciences 11, 9968.
Liu, P., Biljecki, F., 2022. A review of spatially-explicit geoai applications in urban geography. International Journal of Applied Earth Observation and Geoinformation 112, 102936.
Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., Hong, Y., 2017. Classifying urban land use by integrating remote sensing and social media data. International Journal of Geographical Information Science 31, 1675–1696.
Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015. Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers 105, 512–530.
Liu, Y., Zhang, X., Kong, X., Wang, R., Chen, L., 2018. Identifying the relationship between urban land expansion and human activities in the yangtze river economic belt, china. Applied Geography 94, 163–177.
Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S., 2020. 12-in-1: Multi-task vision and language representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10437–10446.
Lu, W., Tao, C., Li, H., Qi, J., Li, Y., 2022. A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. Remote Sensing of Environment 270, 112830.
Lu, X., Wang, B., Zheng, X., Li, X., 2017. Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing 56, 2183–2195.
Murali, N., Shanthi, A., 2022. Remote sensing image captioning via multilevel attention-based visual question answering, in: Innovations in Computational Intelligence and Computer Vision. Springer, pp. 465–475.
Ning, H., Li, Z., Ye, X., Wang, S., Wang, W., Huang, X., 2022. Exploring the vertical dimension of street view image based on deep learning: A case study on lowest floor elevation estimation. International Journal of Geographical Information Science 36, 1317–1342.
Niu, H., Silva, E.A., 2021. Delineating urban functional use from points of interest data with neural network embedding: A case study in greater london. Computers, Environment and Urban Systems 88, 101651.
Pađen, I., García-Sánchez, C., Ledoux, H., 2022. Towards automatic reconstruction of 3D city models tailored for urban flow simulations. Frontiers in Built Environment 8, 899332. doi:10.3389/fbuil.2022.899332.
Qi, Y., Chodron Drolma, S., Zhang, X., Liang, J., Jiang, H., Xu, J., Ni, T., 2020. An investigation of the visual features of urban street vitality using a convolutional neural network. Geo-spatial Information Science 23, 341–351.
Qiao, Z., Yuan, X., 2021. Urban land-use analysis using proximate sensing imagery: a survey. International Journal of Geographical Information Science 35, 2129–2148.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.
Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7008–7024.
von Richthofen, A., Herthogs, P., Kraft, M., Cairns, S., 2022. Semantic city planning systems (scps): A literature review. Journal of Planning Literature , 08854122211068526.
Shen, X., Liu, B., Zhou, Y., Zhao, J., Liu, M., 2020. Remote sensing image captioning via variational autoencoder and reinforcement learning. Knowledge-Based Systems 203, 105920.
Song, Z., Wang, H., Qin, S., Li, X., Yang, Y., Wang, Y., Meng, P., 2022. Building-level urban functional area identification based on multi-attribute aggregated data from cell phones—a method combining multidimensional time series with a som neural network. ISPRS International Journal of Geo-Information 11, 72.
Suryowati, K., Bekti, R., Faradila, A., 2018. A comparison of weights matrices on computation of dengue spatial autocorrelation, in: IOP Conference Series: Materials Science and Engineering, IOP Publishing. p. 012052.
Traag, V.A., Waltman, L., Van Eck, N.J., 2019. From louvain to leiden: guaranteeing well-connected communities. Scientific reports 9, 1–12.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 6000–6010.
Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903 .
Venerandi, A., Feliciotti, A., Fleischmann, M., Kourtit, K., Porta, S., 2022. Urban form character and Airbnb in Amsterdam (NL): A morphometric approach. Environment and Planning B: Urban Analytics and City Science , 23998083221115196doi:10.1177/23998083221115196.
Van de Voorde, T., Jacquet, W., Canters, F., 2011. Mapping form and function in urban areas: An approach based on urban metrics and continuous impervious surface data. Landscape and Urban Planning 102, 143–155.
Wang, J., Cai, X., Chen, F., Zhang, Z., Zhang, Y., Sun, K., Zhang, T., Chen, X., 2020. Hundred-year spatial trajectory of lake coverage changes in response to human activities over wuhan. Environmental Research Letters 15, 094022.

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 of Things Journal .

Wang, P., Zhang, T., Zheng, Y., Hu, T., 2022b. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. International Journal of Geographical Information Science 36, 1231–1257.

Wang, R., Feng, Z., Pearce, J., Yao, Y., Li, X., Liu, Y., 2021. The distribution of greenspace quantity and quality and their association with neighbourhood socioeconomic conditions in guangzhou, china: A new approach using deep learning method and street view images. Sustainable Cities and Society 66, 102664.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2020. Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp. 38–45.

Xu, X., Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., Luo, D., 2022a. Associations between street-view perceptions and housing prices: Subjective vs. objective measures using computer vision and machine learning techniques. Remote Sensing 14, 891.

Xu, Y., Jin, S., Chen, Z., Xie, X., Hu, S., Xie, Z., 2022b. Application of a graph convolutional network with visual and semantic features to classify urban scenes. International Journal of Geographical Information Science , 1–26.

Xu, Y., Zhou, B., Jin, S., Xie, X., Chen, Z., Hu, S., He, N., 2022c. A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method. Computers, Environment and Urban Systems 95, 101807.

Yang, M., Kong, B., Dang, R., Yan, X., 2022a. Classifying urban functional regions by integrating buildings and points-of-interest using a stacking ensemble method. International Journal of Applied Earth Observation and Geoinformation 108, 102753.

Yang, Q., Ni, Z., Ren, P., 2022b. Meta captioning: A meta learning based remote sensing image captioning framework. ISPRS Journal of Photogrammetry and Remote Sensing 186, 190–200.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., Mai, K., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. International Journal of Geographical Information Science 31, 825–848.

Yao, Y., Zhang, J., Qian, C., Wang, Y., Ren, S., Yuan, Z., Guan, Q., 2021. Delineating urban job-housing patterns at a parcel scale with street view imagery. International Journal of Geographical Information Science 35, 1927–1950.

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. International Journal of Applied Earth Observation and Geoinformation 103, 102514.

Yu, B., Lee, Y., Sohn, K., 2020. Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (gcn). Transportation research part C: emerging technologies 114, 189–204.

Yu, W., Ai, T., He, Y., Shao, S., 2017. Spatial co-location pattern mining of facility points-of-interest improved by network neighborhood and distance decay effects. International Journal of Geographical Information Science 31, 280–296.

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

Zhang, F., Wu, L., Zhu, D., Liu, Y., 2019a. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. ISPRS journal of photogrammetry and remote sensing 153, 48–58.

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H., Ratti, C., 2018. Measuring human perceptions of a large-scale urban region using machine learning. Landscape and Urban Planning 180, 148–160.

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

Zhang, S., Tong, H., Xu, J., Maciejewski, R., 2019b. Graph convolutional networks: a comprehensive review. Computational Social Networks 6, 1–23.

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. Building and Environment 198, 107883.

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

Zhang, Y., Zhang, F., Chen, N., 2022a. Migratable urban street scene sensing method based on vision language pre-trained model. International Journal of Applied Earth Observation and Geoinformation 113, 102989. URL: https://www.sciencedirect.com/science/article/pii/S1569843222001807, doi:https://doi.org/10.1016/j.jag.2022.102989.

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

Zhao, B., 2021. A systematic survey of remote sensing image captioning. IEEE Access 9, 154086–154111.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., Li, H., 2019. T-gcn: A temporal graph convolutional network for traffic prediction. IEEE Transactions on Intelligent Transportation Systems 21, 3848–3858.

Zhao, T., Huang, Z., Tu, W., He, B., Cao, R., Cao, J., Li, M., 2022. Coupling graph deep learning and spatial-temporal influence of built environment for short-term bus travel demand prediction. Computers, Environment and Urban Systems 94, 101776. doi:10.1016/j.compenvurbsys.2022.101776.

Zhou, G., Li, C., Zhang, J., 2020. Identification of urban functions enhancement and weakening based on urban land use conversion: A case study of changchun, china. Plos one 15, e0234522.

Zhu, D., Zhang, F., Wang, S., Wang, Y., Cheng, X., Huang, Z., Liu, Y., 2020. Understanding place characteristics in geographic contexts through graph convolutional neural networks. Annals of the American Association of Geographers 110, 408–420.