CureGraph: Contrastive multi-modal graph representation learning for urban living circle health profiling and prediction

Jinlin Li, Xiao Zhou [ID],*

Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

A R T I C L E   I N F O

A B S T R A C T

The early detection and prediction of health status decline among the elderly at the neighborhood level are of great significance for urban planning and public health policymaking. While existing studies affirm the connection between living environments and health outcomes, most rely on single data modalities or simplistic feature concatenation of multi-modal information, limiting their ability to comprehensively profile the health-oriented urban environments. To fill this gap, we propose **CureGraph**, a **c**ontrastive m**u**lti-modal **r**epresentation learning framework for urban h**e**alth prediction that employs **graph**-based techniques to infer the prevalence of common chronic diseases among the elderly within the urban living circles of each neighborhood. CureGraph leverages rich multi-modal information, including photos and textual reviews of residential areas and their surrounding points of interest, to generate urban neighborhood embeddings. By integrating pre-trained visual and textual encoders with graph modeling techniques, CureGraph captures cross-modal spatial dependencies, offering a comprehensive understanding of urban environments tailored to elderly health considerations. Extensive experiments on real-world datasets demonstrate that CureGraph improves the best baseline by 28% on average in terms of R^2 across elderly disease risk prediction tasks. Moreover, the model enables the identification of stage-wise chronic disease progression and supports comparative public health analysis across neighborhoods, offering actionable insights for sustainable urban development and enhanced quality of life. The code is publicly available at https://github.com/jinlin2021/CureGraph.

1. Introduction

Over the past two decades, rapid urbanization and the global trend of an ageing population have emerged as inevitable and intersecting demographic forces shaping the current century. According to statistics [41], no country had more than 18 percent of its population over the age of 65 in 2000. However, due to the accelerating ageing process, this proportion is projected to reach 38 percent by 2050. This demographic shift would inevitably pose a series of significant challenges to cities worldwide, underscoring the need for more age-friendly urban environments. Among the various factors influencing the quality of life for the elderly, chronic conditions that increase with age are critical and must not be overlooked. In 2020, the UN General Assembly declared 2021–2030 as the Decade of Healthy Aging [39], highlighting the urgent need for policymakers worldwide to prioritize initiatives aimed at enhancing the well-being of older individuals. Therefore, devising an effective health-oriented profiling and prediction framework for aging-friendly urban environments—grounded in a deep understanding of the linkages between the multi-modal semantics of

neighborhood living circles and chronic diseases associated with aging—is essential in this era of longevity and forms the core research focus of this paper.

In current literature, a substantial body of research has revealed that the characteristics of human settlements are intricately associated with the chronic health conditions of urban senior citizens. For instance, Cassarino and Setti [3] demonstrated that a community's social and built environment can influence elderly cognition, with complex neighborhood settings potentially delaying cognitive decline. A study on Hong Kong communities [15] found that the relationship between street connectivity and mental health follows an inverted U-shaped pattern, while park green spaces and mixed land use positively impact mental health and subjective well-being. Wang et al. [50] utilized the street view images to assess how neighborhood walkability factors-such as sky visibility, greenery, and building proportion-affect elderly mental health, particularly regarding depression and anxiety. Similarly, Song et al. [43] found that green and blue spaces, population density, and public transport accessibility present statistically significant relationships with blood pressure. Maharana and Nsoesie [34] applied a convolutional neural network (CNN) to extract built environment features from high-resolution satellite images, exploring links between physical attributes (e.g., parks, highways, green streets, pedestrian crossings, and housing types) and the health conditions of community residents.

In addition to the physical environment of neighborhoods, another line of research has focused on the connections between lifestyle choices and residents' health status. For instance, Li et al. [26] examined the prevalence and potential risk factors of mild cognitive impairment (MCI), highlighting that a healthy diet, participation in physical, intellectual, and social activities, and effective management of vascular-related diseases are crucial for maintaining cognitive and brain health. Galiatsatos et al. [11] investigated the relationships between socio-economic determinants, tobacco store density, and health outcomes at the neighborhood level. Similarly, Hasthanasombat and Mascolo [16] analyzed the causal impact of sports venue availability on the prevalence of antidepressant prescriptions in London neighborhoods. Wang et al. [53] utilized check-in data from various points of interest (POIs) to capture local lifestyle patterns within neighborhoods and predict the progression rates of chronic diseases.

Although there is a multitude of studies advocate for actions on the built environment and lifestyle choices to improve health outcomes for the elderly, the evidence base for these findings often relies on urban data from a single modality (e.g., social demography, street view images, or textual reviews on social networks) to explore relationships between specific neighborhood characteristics and particular diseases. A health-focused neighborhood representation learning framework that comprehensively considers the scope of the elderly's daily activities and leverages rich multi-modal urban data to predict age-related health declines is still lacking. To address this gap, this study proposes a contrastive multi-modal graph representation learning approach for profiling the health of the elderly in urban communities. The novelty of this research lies in three key aspects. First, from the perspective of *spatial scope*, it offers a contextualized view on healthy aging-focused built environment studies by jointly analyzing both intra-community and inter-community conditions. Second, in terms of *methodology*, it utilizes multi-modal urban data and introduces fusion techniques for hierarchical neighborhood representation learning. Finally, in *practical terms*, it provides actionable recommendations for public health practitioners and urban planners to enhance spatial planning for healthy aging.

More specifically, since neighborhoods serve as the primary venue for the elderly to carry out their daily activities and are a crucial determinant of their quality of life—particularly due to declining physical functions, social adaptability, and increased environmental vulnerability [55,56,1]—we introduce the concept of the *urban community living circle* [37,33] as the core unit of study. In this framework, the 15-minute living circle surrounding each residential area is considered the main activity space for elderly residents. This circle not only encompasses the physical elements within the residential district itself but also includes surrounding facilities that can be accessed within a 15-minute walk or bike ride. This approach enables us to account for both the physical environment of the neighborhood and the lifestyles shaped by nearby amenities. To fully capture the scope of the community living circles, we collect multi-modal datasets, including textual descriptions, reviews, and images of the target residential area and the POIs within each living circle. For better illustration, we present the distribution of the multi-modal data within the 15-minute community living circle in Beijing, as shown in Fig. 1.

We leverage pre-trained models to design contrastive learning tasks, utilizing a combination of geographic distance and data augmentation techniques. First, we use CNN-based methods and contrastive learning tasks based on geographic distance to extract visual features. Next, we introduce a cross-modal contrastive learning task to learn descriptive features from long-text descriptions within the living circle. Simultaneously, we apply supervised contrastive learning to extract review features from POIs, incorporating sentiment analysis as an additional layer of comparison. The features learned across modalities are then mapped to a common space via the modal encoder. To account for the spatial dependencies between urban areas, we propose an empirical measure—namely, the spatial autocorrelation matrix—to capture the geographical relationships and similarities in POI distribution between community living circles. By adopting this strategy, our framework can simultaneously integrate multi-modal data from the living circles while preserving local spatial constraints. Finally, through Graph Convolutional Network (GCN) encoding, with residual links and node splicing across modalities, we generate a representation for each living circle to predict the prevalence of chronic diseases among the elderly. The resulting graph embeddings also enable comparative health analyses across urban regions at multiple levels.

The contributions of this work are summarized as follows:

- To the best of our knowledge, this is the first study to leverage multi-modal information with local constraints from urban living circles to address health issues among the elderly.
- We propose a novel contrastive multi-modal graph-based model for urban area representation learning, capturing both inter-modality interactions across different living circles and intra-modality interactions within each circle.
- By introducing a spatial autocorrelation matrix for community living circles, our approach provides new insights into utilizing the spatial characteristics of living circle elements to predict aging-related chronic diseases.
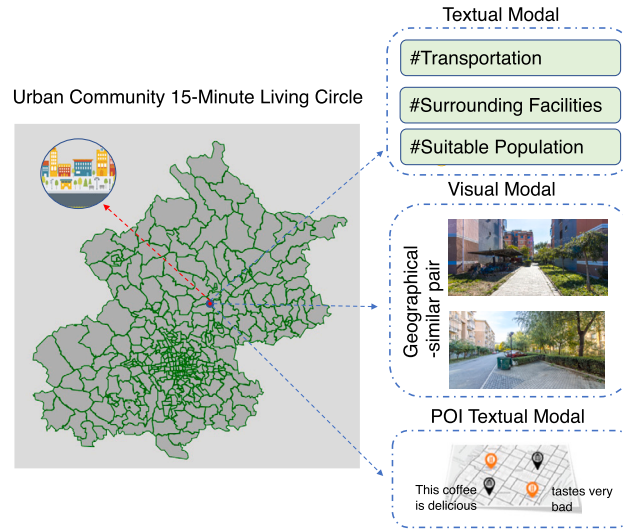
Fig. 1. Multi-modal data of urban community 15-minute living circles.

- The proposed framework is highly flexible and versatile, capable of uncovering urban health disparities across multiple dimensions. It also facilitates the integration of additional neighborhood features (e.g., housing prices, satellite imagery) to enrich embedding learning.
- We validate the effectiveness of the proposed framework using real-world urban datasets. Ablation studies further confirm the benefits of incorporating multi-modal information for neighborhood representation learning, contributing to healthy, age-friendly community planning.

2. Related work

2.1. Contrastive learning in urban settings

Inspired by significant advancements in self-supervised learning within computer vision [4,17,52] and natural language processing [57,12,60], researchers have begun to extend these methodologies to urban modal coding. Despite these innovations, much of the existing research in this domain remains grounded in Tobler's First Law of Geography [36] when designing contrastive learning similarity measures. Self-supervised learning, particularly contrastive learning, has been widely used to distill expressive modal elements in urban spaces. For example, Jean et al. [23] employed triplet loss to minimize the distance between representations of spatially proximate satellite images while maximizing the distance for spatially distant pairs. Similarly, Wang et al. [54] utilized a comparable loss function for street view imagery. Li et al. [25] applied the InfoNCE loss [38] from the SimCLR framework [4] to extract visual representations from satellite and street view images. Xi et al. [58] expanded this approach by incorporating a POI-based similarity metric, ensuring that images associated with similar POI category distributions yielded closer visual representations. Liu et al. [30] further extended this line of research by integrating a knowledge graph (KG) to represent urban knowledge within satellite imagery, formulating an image-KG contrastive loss that builds upon the traditional InfoNCE loss [38].

However, the aforementioned studies rely solely on contrastive learning frameworks based on single-modality data or perspectives, overlooking the complementary potential of combining urban images and textual information.

2.2. Urban region representation learning

Urban region representation learning has garnered significant attention in recent years, driven by the proliferation of urban big data, the rapid advancements in embedding techniques within deep learning, and the pressing need for applications in smart cities. Broadly, research in this area often leverages word embedding or graph embedding techniques, tailored to address the spatio-temporal characteristics unique to urban settings.

For example, Wang and Li [49] constructed a mobility graph using fine-grained taxi commuting data and learned region embeddings at different time intervals. Similarly, Yao et al. [62] used taxi trajectories to learn embeddings of urban functional zones, applying the word2vec [35] method by treating each urban region as a word and mobility events as context. More recently, the urban computing community has increasingly recognized the importance of exploiting multi-modal urban data for region representation learning, given the complexity of cities as systems. A notable example is Urban2Vec by Wang et al. [54], a neighborhood embedding method based on word embedding techniques that integrates multi-modal data, including street view images and POI textual reviews. From a multi-view perspective, Fu et al. [10] incorporated spatial relationships between regions, constructing multi-view POI-POI networks to represent urban regions. Zhang et al. [64] advanced this approach by designing a cross-view information-sharing strategy

and a weighted multi-view fusion technique to combine human mobility and POI data. Similarly, Huang et al. [21] utilized multi-modal data as node and edge features within a multi-graph, leveraging human mobility data to represent neighborhood relationships while employing contrastive sampling to learn neighborhood representations. Li et al. [27] proposed RegionDCL, an unsupervised framework utilizing OpenStreetMap (OSM) building footprints and POI data to learn regional representations. This approach incorporates dual contrastive learning at both the group and region levels, preserving spatial proximity and architectural similarity among building clusters. Yong and Zhou [63] proposed the MuseCL framework, which leverages contrastive learning across visual and textual modalities, combined with a cross-modality attentional fusion module, to enable fine-grained urban region profiling and socioeconomic prediction. Xu and Zhou [59] introduced the CGAP algorithm, incorporating hierarchical graph pooling with local and global attention mechanisms to improve urban region representation by integrating multi-modal data such as POIs and inter-regional human mobility patterns.

Given that graphs are well-suited for organizing urban elements and their complex relationships, and graph neural network (GNN)-based approaches have recently emerged as powerful tools for graph embedding tasks [48,13], this study adopts graph convolutional networks (GCNs) as the base model to encode multi-modal graphs encompassing diverse urban data modalities. Additionally, our framework integrates advanced techniques, including contrastive learning, image and text analysis, and spatial pattern mining. This combination of methods distinguishes our community living circle representation learning framework from existing approaches, offering a novel and comprehensive solution.

2.3. Health in urban environments

As the introduction section provides a comprehensive overview of studies related to health in urban built environments, we will not repeat them here. A key limitation in the current literature is the frequent neglect of non-linear interactions among urban features, which can significantly influence community health outcomes. Many conventional studies that employed traditional spatial statistics methods [28,34,50,46] assumed linear relationships between features, overlooking the inherent complexity and subtle dynamics of urban environments. In reality, urban health outcomes are influenced not only by the geographic distribution of facilities and resources, as described by Tobler's First Law of Geography, but also by the complex interactions among various neighborhood characteristics, including social [20,66], economic, mobility [65], and environmental factors. This complexity is consistent with the Second Law of Geography, which highlights spatial heterogeneity [2].

To address these limitations, our proposed method captures the relationships between feature modalities within urban living circles while accounting for spatial heterogeneity across different circles. By leveraging advanced techniques such as contrastive learning, graph representation learning, and spatial analysis, our approach provides a more nuanced understanding of the intricate interplay between factors influencing urban health outcomes at multiple scales. Additionally, it offers valuable insights for urban planners and policymakers, guiding strategies to improve the health and well-being of elderly populations in urban communities.

3. Problem statement

In this section, we delineate the related notations and the problem we study.

Definition 1 *(Urban living circle).* The health condition of a metropolitan area can be reflected by a series of urban living circle scenes. Among these, the community living circle serves as a fundamental spatial unit capable of meeting the essential daily activity requirements of its residents. These activities typically have high frequency, short duration, and are confined to the residential area and its immediate vicinity. In this study, we define the urban living circle as the service radius of a residential area, specifically referring to a 15-minute walking distance (within a 1000-meter radius) from various essential service facilities. This concept is commonly referred to as the "15-minute living circle," which also serves as the basic spatial unit for our downstream community health predictions. Given a set of urban living circles, $C = \{c_1, c_2, \ldots, c_n\}$, each urban living circle c_i contains a paragraph of residential area text, $\mathcal{T}_i$, a set of residential area images, $\mathcal{V}_i = \{v_{i1}, v_{i2}, \ldots, v_{iM_i}\}$, and a set of health-related POIs, $\mathcal{P}_i = \{p_{i1}, p_{i2}, \ldots, p_{iO_i}\}$ located within c_i. Here, n represents the number of residential areas, and M_i and O_i denote the number of images and POIs contained within each living circle c_i, respectively. We assume that the GPS coordinates of the images and the residential area are consistent.

Definition 2 *(Urban living circle texts).* The living circle text describes the supporting facilities, transportation options, surrounding amenities, and the target residents of the residential area. The representation of urban living circles texts, denoted as $\mathcal{T}$, is formalized as follows:

$$\mathcal{T} = \left\{ \vec{c}_1^{\,t}, \vec{c}_2^{\,t}, \ldots, \vec{c}_i^{\,t} \right\}, \vec{c} \in \mathbb{R}^{F_t}, \forall \vec{c^t} \in \mathcal{T}, \tag{1}$$

where c_i^t is the corresponding textual feature of i-th region and F_t is the number of dimensions of the feature.

Definition 3 *(Urban living circle images).* The visual modality of the living circle captures the physical environment within the residential area and serves as a visual complement to the textual modality. The number of community images in each living circle may vary. We denote the set of urban living circle images as $\mathcal{V}$, defined as follows:

$$\mathcal{V} = \left\{ \vec{c}_1^{\,v}, \vec{c}_2^{\,v}, \ldots, \vec{c}_i^{\,v} \right\}, \vec{c} \in \mathbb{R}^{F_v}, \forall \vec{c^v} \in \mathcal{V}, \tag{2}$$

Table 1
Descriptions of frequently used notations.

Notation	Description
c_i^v	The corresponding visual feature of living circle c_i.
c_i^t	The corresponding textual feature of living circle c_i.
c_i^P	The corresponding POI textual feature of living circle c_i.
$\text{sim}(\cdot)$	The cosine similarity.
M_i	Number of images contained within each living circle c_i.
e_i	$e_i \in \mathbb{R}^d$ is the embedding of c_i.
τ	$\tau \in R$ is the temperature hyperparameter in modal encoder.
$S_{i,j}$	The spatial autocorrelation coefficients between c_i and c_j.
C_i^{Top-K}	The top-k most correlated candidate regions for living circle c_i.

where c_i^v is the corresponding visual feature of i-th region and F_v is the number of dimensions of the feature.

Definition 4 *(Urban living circle POI texts).* POIs serve as direct representations of urban functions. The distinctive features of check-ins and review information associated with different types of POIs can be leveraged to infer the healthiness of lifestyles. We map the relevant POIs to their corresponding living circle locations and denote the set of urban living circle POI texts as $\mathcal{P}$, defined as follows:

$$\mathcal{P} = \left\{ \vec{c}_1^p, \vec{c}_2^p, \dots, \vec{c}_i^p \right\}, \vec{c} \in \mathbb{R}^{F_p}, \forall \vec{c^p} \in \mathcal{P}, \tag{3}$$

where c_i^p is the corresponding POI textual feature of the i-th region, and F_p is the dimensionality of the feature. Each POI is represented by textual data, including texts extracted from its categories and check-in reviews.

Definition 5 *(Spatial correlation).* The spatial correlation of a living circle is represented by the spatial autocorrelation matrix. This matrix is a variant of the covariance matrix of regions, capturing the pairwise similarity between each pair of 15-minute living circles.

$$\mathcal{H} = \{\vec{h}_i \mid i = 1, 2, \dots, n\}, \vec{h}_i = (c_1, c_2, \dots, c_{|\vec{h}_i|}), \tag{4}$$

where $\vec{h}_i$ is a spatial neighborhood vector for an urban living circle c_i.

Problem Statement. Given a set of urban living circle scenes, C, along with the corresponding residential area texts $\mathcal{T}$, visual data $\mathcal{V}$, and POI textual data $\mathcal{P}$, the main objective of the proposed model is to learn a distributed, low-dimensional health embedding, $\vec{e}$, for each urban living circle. Using these learned health embeddings as input, the model aims to predict the prevalence of four common geriatric diseases among the elderly within each urban living circle, namely Mild Cognitive Impairment (MCI), hypertension, diabetes, and Major Depressive Disorder (MDD). Additionally, the model seeks to generate health profiles at various spatial levels.

For brevity, frequently used notations are listed in Table 1. The health embeddings for each urban living circle are represented as:

$$\mathcal{E} = \left\{ \vec{e}_1, \vec{e}_2, \dots, \vec{e}_n \right\}, \vec{e} \in \mathbb{R}^d, \forall \vec{e} \in \mathcal{E}, \tag{5}$$

where $\vec{e}_i \in \mathbb{R}^d$ represents the health embedding of the i-th urban living circle c_i, and d is the embedding size.

4. Methodology

In this section, we present the framework for graph representation learning based on multimodal spatial fusion to generate health embeddings for urban living circles. The proposed framework integrates visual and textual data from residential areas alongside corresponding POI information within each living circle to learn a compact, low-dimensional, and spatially-aware vector representation. This representation is then used to predict the prevalence of four common geriatric diseases among the elderly population in urban areas. An overview of the proposed CureGraph framework is illustrated in Fig. 2.

4.1. Modality encoder

To accurately extract features from urban multimodal data, we draw inspiration from the success of contrastive learning in modality encoding. We develop multiple types of contrastive learning encoders to effectively extract features from textual, visual, and POI data related to urban life in residential areas. The core principle of contrastive learning is to achieve both alignment and uniformity [51]. Specifically, contrastive learning assumes that samples from positive pairs are similar, aiming to maximize the similarity of representations within the same pair while increasing the dissimilarity between representations of different pairs. Consequently, constructing contrastive pairs is a critical step in this approach. In our modality encoder, we leverage both self-similarity and geographical similarity across modalities to construct contrastive pairs. The encoder is designed to map similar data to closely aligned embeddings while maximizing the divergence between embeddings of dissimilar data types.
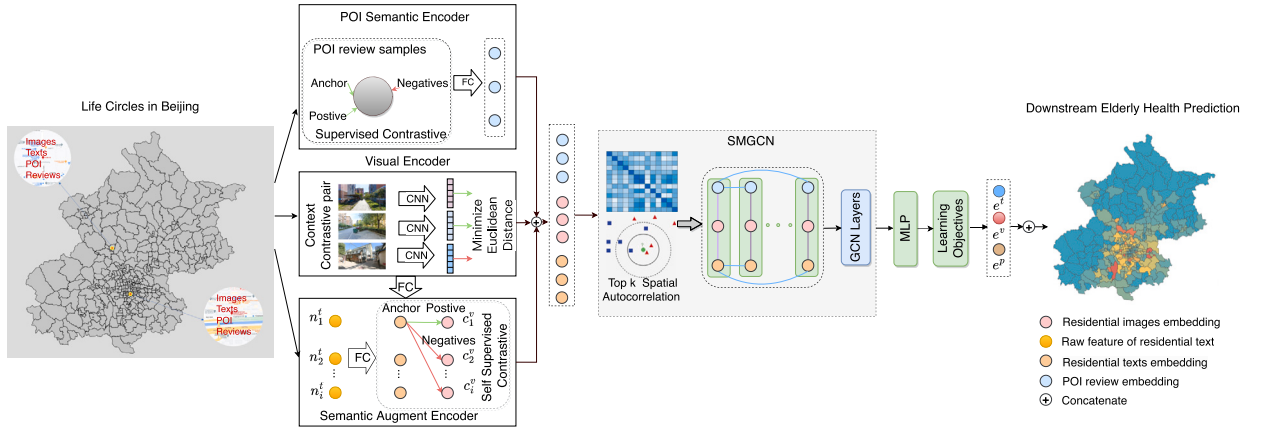
Fig. 2. Framework overview of the proposed CureGraph. The visual encoder leverages self-supervised contrastive learning with negative sampling to extract image features. The POI semantic encoder uses supervised contrastive learning to map POI review texts with the same sentiment label closer in the feature space. Additionally, the semantic augment encoder employs cross-modal contrastive learning techniques, integrating visual features to enrich textual semantics. The spatial multimodal graph network (SMGCN) is constructed by incorporating both modality similarity and spatial correlation. Finally, the multimodal embeddings are processed through a multilayer perceptron (MLP) for downstream elderly health prediction tasks. This framework aims to provide robust health predictions by effectively capturing multimodal and spatial interactions.

4.1.1. Visual modality encoder

Tobler's First Law of Geography states that objects closer to each other are more likely to be related than those farther apart. Building on this principle, we design a geospatial contrastive learning model to effectively extract image features of urban living circles from residential images. The visual contrastive learning model aligns similar samples by positioning them closer in the projection space while pushing dissimilar samples farther apart. In this context, we assume that images from the same residential area share similar geospatial information.

To capture this, we define the geographic neighbors of a residential area as its "context", determined by the service radius of the living circle. Specifically, for an image v, we identify a positive image v_p within the context (geospatial neighbors) and a negative image v_n outside the context. The model is trained on a collection of triplets (v, v_p, v_n) using a geospatial contrastive loss function:

$$\mathcal{L}_v^{geo}\left(v, v_p, v_n\right) = \left[m + \|\boldsymbol{x} - \boldsymbol{x_p}\|_2 - \|\boldsymbol{x} - \boldsymbol{x_n}\|_2\right]_+, \tag{6}$$

where $[\cdot]_+$ is the rectifier function, and $\|\cdot\|_2$ denotes the Euclidean distance. The margin m ensures that the difference between these distances does not grow infinitely large. The embeddings of images v, v_p, and v_n are represented as $\boldsymbol{x}$, $\boldsymbol{x_p}$, and $\boldsymbol{x_n}$, respectively. These embeddings are derived by encoding the images using a pre-trained InceptionV3 [44] model with an embedding dimension of 768.

We also design a contrastive learning framework based on data augmentation to enhance the robustness of feature extraction. The principle of *self-similarity* ensures that augmented data variants remain similar to the original samples. In computer vision [22], contrastive samples of an image can be generated through data augmentation techniques such as rotation, cropping, masking, and adding noise. For this study, we use rotation and flipping to create two augmented images, v and v^+, for each original image. Let $\boldsymbol{x}$ and $\boldsymbol{x}^+$ represent their embeddings, and consider other images in the same batch as negative samples. The loss function for a mini-batch of N pairs is expressed as:

$$\mathcal{L}_v^{aug} = -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{x}_{v,i}, \boldsymbol{x}_{v,i}^+\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(\boldsymbol{x}_{v,i}, \boldsymbol{x}_{v,j}^+\right)/\tau\right)}, \tag{7}$$

where $\text{sim}(\cdot)$ denotes the cosine similarity between two vectors, and $\tau \in \mathbb{R}$ is a temperature hyperparameter that controls the concentration of the distribution.

Finally, the overall contrastive loss for the visual modality encoder, $f_v(\cdot)$, combines geospatial and augmentation-based components:

$$\mathcal{L}_v = \mathcal{L}_v^{geo} + \mathcal{L}_v^{aug}. \tag{8}$$

After obtaining the feature representation for each image through the encoder $f_v(\cdot)$, we apply an averaging operation to derive the corresponding visual feature for each urban living circle c_i. This is expressed as:

$$\boldsymbol{c}_i^v = \frac{1}{M_i} \sum_{j=1}^{M_i} \boldsymbol{x}_{v,ij}, \tag{9}$$

where M_i denotes the number of images in the living circle c_i, and $\boldsymbol{x}_{v,ij}$ represents the raw feature representation of the j-th image extracted by the visual modality encoder.

4.1.2. Textual augmentation encoder

In the field of computer vision (CV), contrastive learning has proven to be a powerful method for feature extraction. Similarly, in natural language processing (NLP), contrastive learning is gaining traction due to its ability to learn robust representations using contrastive pairs that pull similar samples closer while pushing dissimilar ones apart [51]. A critical aspect of contrastive learning lies in constructing positive and negative samples, with various text augmentation techniques—such as adversarial attacks, sentence shuffling, cutoff, and dropout—being widely adopted [60].

Building on this principle, we leverage geographical similarity between residential texts and images that share the same spatial information. Our fundamental assumption is that texts and images in close geographical proximity exhibit similar semantic spatial representations. Notably, textual descriptions of residential areas within living circles often consist of long sentences. To capture the full semantic modality of text, we draw inspiration from cross-modal contrastive learning, aligning textual representations with their visual counterparts.

For the text modality encoder, we introduce a visually-enhanced text semantic augmentation model based on cross-modal contrastive learning. The core insight is that the semantic representation of a region should closely align with its visual features. Specifically, we use the pre-trained language model `hfl/chinese-bert-wwm-ext`, following [6], to encode the textual modality. A fully connected network then projects the representations of both textual and visual modalities into a shared latent space for contrastive learning. The outputs for textual and visual modalities of $c_i \in C$ are expressed as:

$$
\begin{aligned}
h_{t,i} &= W_e^t \boldsymbol{n}_i^t + b_i^t, \\
h_{v,i} &= W_e^v \boldsymbol{c}_i^v + b_i^v,
\end{aligned}
\tag{10}
$$

where $\boldsymbol{n}_i^t$ represents the raw feature vector of the textual modality, and $\boldsymbol{c}_i^v$ is the visual embedding of living circle c_i. $h_{t,i}$ and $h_{v,i}$ denote the hidden features of the textual and visual modalities, respectively.

To align textual and visual modalities, we extend the traditional InfoNCE loss for the textual augmentation encoder. The contrastive loss is defined as:

$$
\mathcal{L}_t^{aug} = -\log \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{h}_{t,i}, \boldsymbol{h}_{v,i}\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{sim}\left(\boldsymbol{h}_{t,i}, \boldsymbol{h}_{v,j}\right)/\tau\right)},
\tag{11}
$$

where the loss is computed over a mini-batch of N samples.

Unlike previous studies on urban region representation that rely on single or multi-view-based contrastive loss within the same modality or employ standard data augmentation techniques, our proposed cross-modal contrastive loss integrates visual information captured in living circles into textual representations. This innovative approach leverages the complementary nature of visual and textual data, enhancing the overall understanding of urban regions and enabling more accurate and comprehensive representations.

4.1.3. POI textual modality encoder

We employ a combination of geographic similarity and data augmentation techniques to train the POI textual modality encoder. Unlike the unsupervised learning approach in the textual augmentation encoder, this encoder utilizes a supervised contrastive learning model. To ensure semantic accuracy and effectively evaluate the quality of services within a community, we incorporate the sentiment tone and category of POI review texts within the living circle. The primary objective of the POI textual modality encoder is to group reviews with similar sentiment tones while distinguishing those with differing sentiments.

Specifically, we map relevant POIs to the designated living circle, guided by the service radius of the residential area. In order to ensure proximity among POI review text embeddings within the same sentiment category, we employ POI ratings as labels to categorize review texts accordingly. For each POI review text p located within the living circle, we utilize the dropout technique twice as a method of data augmentation. This strategy aims to enrich the dataset with additional positive examples, thereby facilitating the generation of a greater number of positive sample pairs. In the field of NLP, the essence of contrastive learning techniques lies in the creation of positive pairs characterized by sufficient variability. This variability is crucial for training models to discern distinct text representations. Unlike traditional data augmentation methods, which directly alter text content (e.g., through cropping or word deletion) and may inadvertently modify the original intent or introduce noise, our approach leverages the BERT model. By applying diverse dropout masks twice, we generate two unique representations for the same sentence without altering the text directly. This method not only maintains the original text's integrity but also exploits the model's inherent randomness to enhance the diversity between representations.

The goal of our model is to increase the similarity scores among samples within the same category while reducing similarity scores for samples from different categories. Let $\boldsymbol{x}_{p,i}$ and $\tilde{\boldsymbol{x}}_{p,i}$ represent the textual modality embeddings of POI reviews p and $\tilde{p}$, respectively, where $\tilde{p}$ shares the same label as p. The sentiment-based, geographically supervised contrastive loss is computed as:

$$
\mathcal{L}_p = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathrm{sim}(\boldsymbol{x}_{p,i}, \tilde{\boldsymbol{x}}_{p,i})/\tau)}{\sum_{j \in A(i)} \exp(\mathrm{sim}(\boldsymbol{x}_{p,i}, \tilde{\boldsymbol{x}}_{p,j})/\tau)},
\tag{12}
$$

where $P(i) \equiv \left\{ p \in A(i) : \tilde{y}_p = \tilde{y}_i \right\}$ represents the indices of all positive samples in the mini-batch distinct from i, and $|P(i)|$ is its cardinality. $A(i) \equiv I \setminus i$ denotes all indices except i. The textual raw feature $\boldsymbol{x}_p$ is extracted using the pre-trained language model `hfl/chinese-bert-wwm-ext`, as described in [6], with a feature dimension of 768.

To represent each POI within a living circle, we extract semantic features from the review texts and calculate the average sentiment embedding. Given that POI categories are highly correlated with the facility distribution within a living circle, we use the same pre-trained language model to encode the POI categories. The two components—POI review embeddings and category embeddings—are concatenated to form the final POI textual feature. Subsequently, a fully connected network projects the POI textual modality features into the shared latent space for graph learning. Finally, we compute the aggregated POI textual representation c_i^p for living circle c_i as:

$$c_i^p = \sum_{i=1}^{G_i} \sum_{j=1}^{m_j} \frac{1}{m_j} \boldsymbol{x}_{p,ij},$$

$$h_{p,i} = W_e^p \boldsymbol{c}_i^p + b_i^p,$$

(13)

where G_i is the number of POI categories within each living circle, m_j is the number of comments for each POI within the living circle, and $\boldsymbol{x}_{p,ij} = [\boldsymbol{x}_{p,ij}^{rv}, \boldsymbol{x}_{p,ij}^{cg}]$. Here, $\boldsymbol{x}_{p,ij}^{rv}$ is the raw feature representation of the POI reviews extracted by the POI textual modality encoder, and $\boldsymbol{x}_{p,ij}^{cg}$ is the category representation of the corresponding POI.

4.2. Extracting spatial autocorrelation

Urban data are often influenced by geospatial interactions and diffusion, leading to potential correlations and dependencies among data points. This spatial correlation, usually represented by the spatial weight matrix, quantifies the degree of interdependence between data at one location and data at other locations. In this study, spatial autocorrelation coefficients capture both geographical distance and POI-related information within the community living circles. The spatial autocorrelation matrix $\mathcal{H}$ is used to measure the pairwise similarity between each pair of living circles. By incorporating spatial autocorrelation into the graph network, we enhance the learning of modal representations within the community.

To calculate the spatial autocorrelation, we first determine the normalized geographical distance $D_{i,j}$ between the i-th living circle c_i and the j-th living circle c_j. Next, we use the TF-IDF model to assess the importance of each POI category within a living circle, treating POIs as words and a region as a document that describes the functional distribution of the living circle. The urban function similarity between circles is then computed as $F_{i,j} = \text{sim}(c_i, c_j)$.

Finally, the combined spatial autocorrelation coefficient is expressed as:

$$S_{i,j} = \frac{\text{sim}(c_i, c_j)}{\log(D_{i,j} + 1)},$$

(14)

where $S_{i,j}$ denotes the spatial autocorrelation coefficient between c_i and c_j. This approach enables the proposed model to learn embeddings for multi-modal data in living circles while accounting for inter-region autocorrelations.

4.3. Graph construction

In this section, we present our approach to modeling and constructing the relationships between nodes in community living circles. As defined in Section 3, we characterize urban community living circles using a 1000-meter service radius around residential areas. Each community living circle encompasses information such as texts, images, and POIs related to the corresponding residential areas. Based on the modal encoder, we can obtain representations of these multimodal data within each community living circle. Specifically, at the spatial street level, we assume that the streets with n residential areas can be represented as an undirected graph $G = (V, E, N)$, where $V = \left\{ v_i \right\}_{i=1}^{3n}$ denotes the three feature nodes in the urban living circle, E denotes the edges connecting nodes, and N represents the neighborhood of each node. We construct the graph as follows:

- **Nodes.** Each living circle encompasses a residential area, represented by three nodes initialized with the textual modality $\boldsymbol{c}^t$, visual modality $\boldsymbol{c}^v$, and POI textual modality $\boldsymbol{c}^p$, respectively. In accordance with the administrative division at the street level, we construct a graph with $3n$ nodes based on the n living circles in the street.
- **Edges and Edge Weighting.** Given the high spatial autocorrelation within the modal information of community living circles, we assume that each node connects to others within the same community living circle. For example, $\boldsymbol{c}^t$ will be connected to both $\boldsymbol{c}^v$ and $\boldsymbol{c}^p$ within the graph. Additionally, for any two nodes of the same modality in different living circles, we connect the top-k nearest nodes based on geographic proximity and modal similarity. Consequently, the graph contains two types of edges: one type connects nodes of different modalities within the same living circle (intra-community edges), and the other connects nodes of the same modality (inter-community edges). We apply different edge-weighting strategies for these two types of edges.

The edge weight increases as node similarity increases, reflecting the stronger information interaction between similar nodes. For intra-community edges, we capture the similarity between node representations. Following the approach in [42], we use angular similarity to calculate the edge weight between nodes connected in this manner. The edge weight is computed as:

$$W_{ij}^1 = 1 - \frac{\arccos\left(\text{sim}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right)}{\pi},$$

(15)

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the feature representations of the i-th and j-th nodes, and $\text{sim}(\cdot)$ denotes cosine similarity.

For the second type of edge, we adhere to the spatial autocorrelation framework described in Subsection 4.2, employing a top-k locality principle. This ensures that for each urban living circle c_i, we connect only the top-k nearest modal nodes within the living circle. The edge weights are determined by both the geographic proximity and modality similarity of the corresponding residential areas. Nodes that are closer in space within a living circle are assigned higher edge weights, ensuring stronger connections for geographically proximate nodes. The edge weight is calculated as:

$$W_{ij}^2 = \frac{W_{ij}^1}{\log(D_{i,j} + 1)}, \quad c_j \in C_i^{Top-K}, \tag{16}$$

where C_i^{Top-K} represents the top-k most correlated candidate regions for the living circle c_i based on spatial autocorrelation, and $D_{i,j}$ is the Euclidean distance between nodes i and j. In CureGraph, we combine the top-k locality concept with modality representation similarity to optimize the calculation and search for parameters.

Our proposed composition method simulates the intra-region connections within the living circle modals while comprehensively capturing the inter-region correlations in the modality space. This innovative approach enables a more comprehensive understanding of the relationships and dependencies among different modalities, resulting in more accurate and robust representations of urban regions. By incorporating both intra-region and inter-region correlations, we can effectively capture the fundamental structure and inherent patterns in the data, thus boosting the overall performance of our model.

4.4. Base model

We employ a deep GCN as the base model to develop SMGCN, which learns efficient node representations in a graph. By modeling the spatial relationships in a graph's multimodal context, SMGCN captures local structural information and the relationships between nodes, enabling it to generate feature representations for each node. Additionally, SMGCN further enhances these representations by integrating information from neighborhoods across different modalities, improving the understanding of graph structure and inter-node relationships while strengthening the encoding of contextual dependencies.

Given an undirected graph $G = (V, E, N)$ and the input vertex features $[\mathcal{T}, \mathcal{V}, \mathcal{P}]$, we approximate first-order spectral graph convolutions by propagating input data through augmented adjacency and degree matrices. To learn distributed representations of each node in the graph, we adopt a layer-wise propagation rule with residual connections, as proposed in [4]. This update rule for node features is expressed as:

$$H^{(k)} = \sigma \left(\left((1-\alpha)\tilde{\mathcal{P}} H^{(k-1)} + \alpha H^{(0)} \right) \left((1 - \beta_{k-1}) I + \beta_{k-1} W^{(k-1)} \right) \right),$$
$$\beta_k = \log \left(\frac{\eta}{k} + 1 \right), \tag{17}$$

where $H^{(k)}$ denotes the node feature after the k-th aggregation iteration, and $\tilde{\mathcal{P}} = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2}$ is the renormalized graph Laplacian matrix of G. The parameters α, β_k, and η are hyperparameters, while $\sigma(\cdot)$ is the activation function (sigmoid), and $W^{(k)}$ is a learnable weight matrix. The parameter β_k ensures that the weight matrix decays adaptively as more layers are stacked. The initial node features, $H^{(0)}$, are initialized with $[\tilde{h}_{t,i}, h_{v,i}, h_{p,i}]$, where $\tilde{h}_{t,i}$ is the augmented textual feature from the textual augmentation encoder, and I is the identity mapping matrix.

After k aggregation iterations, we obtain the fused node features for different modalities, denoted as et, ev, and e_p. These features are then processed by a multi-layer MLP with ReLU activation to transform the concatenated features corresponding to the same living circle. The final representation is used to predict the prevalence of four common geriatric diseases among the elderly.

The training algorithm for our base model is detailed in Algorithm 1. The final node features are computed as:

$$e = f_\theta(e^*), \tag{18}$$

where $e^* = [e_t, e_v, e_p]$, and $f_\theta(\cdot)$ is an MLP with four layers, parameterized by θ.

4.5. Learning objectives

Given a multi-modal spatial GCN, the features of each node in the graph are updated through k-layer iterations. The primary objective of CureGraph is to learn effective low-rank vectors in the embedding space after multi-modal fusion, while preserving spatial autocorrelation. Specifically, CureGraph aims to achieve two goals: (1) capture the rich and meaningful semantics within living areas and fuse them effectively, and (2) accurately model the spatial autocorrelation between living circles. Furthermore, CureGraph is based on the assumption that the texts within residential areas should be highly correlated with the street-view and POI information in their spatial context. Therefore, the model aims to minimize the distance between neighborhoods that have both street-view and POI representations in the embedding space. To train the model effectively, we design a task that reconstructs the spatial autocorrelation coefficients based on the corresponding embeddings and minimizes their Euclidean distance. The learning objective is defined as follows:

$$\mathcal{L} = \sum_{i,j} \left(S_{i,j} - e_i^T e_j \right)^2 + \lambda \sum_{m \in \{v,p\}} \text{dist}(e_t, e_m), \tag{19}$$

Algorithm 1: Training Algorithm of SMGCN.

Input:

Multimodal embeddings of the training urban living circles:

$\mathcal{T}, \mathcal{V}, \mathcal{P} = \{\bar{c}_1^{t,v,p}, \bar{c}_2^{t,v,p}, \ldots, \bar{c}_i^{t,v,p}\}$

Spatial correlation of the training urban living circles:

$\mathcal{H} = \{\vec{h}_i \mid i = 1, 2, \ldots, n\}, \vec{h}_i = (c_1, c_2, \ldots, c_{|\vec{h}_i|})$ and c_i is the i-th living circle, C_i^{Top-K} is the top-k most correlated candidate regions for c_i based on $\mathcal{H}$.

Output:

Learned distributed and low dimensional health embeddings e for urban living circle.

1 **for** $t \leftarrow 1$ *to Epoch* **do**

2     $h_{v,i} = Linear(c_i^v), h_{p,i} = Linear(c_i^p)$;

3     $W_{ij}^1 \leftarrow 1 - \frac{\arccos(\text{sim}(x_i, x_j))}{\pi}$, $x \in \{\mathcal{T}, \mathcal{V}, \mathcal{P}\}$;

4     $W_{ij}^2 \leftarrow \frac{W_{ij}^1}{\log(D_{i,j}+1)}$, $c_j \in C_i^{Top-K}$;

5   conduct graph $G = (V, E, N)$, initialized with $[\vec{h}_{t,i}, h_{v,i}, h_{p,i}]$;

6     $e^* = f([e_t, e_v, e_p])$ where f is the base model ;

7     $e = f_\theta(e^*)$;

8     $loss \Longleftarrow \mathcal{L}$ where $\mathcal{L}$ is Equation (19) ;

9   optimizer.step($loss$) ;

10 **end**

---

where $S_{i,j}$ represents the spatial autocorrelation coefficient between living circles c_i and c_j, and dist($\cdot$) is the distance function used in the vector space (here, we use Euclidean distance). The hyperparameter λ controls the contribution of the spatial distance loss to the global loss function. The term e_m, for $m \in \{v, p\}$, refers to the visual and POI textual modality embeddings, respectively.

5. Experiments

In this section, we conduct experiments on real-world datasets to evaluate the performance of our method in comparison to various baselines. Additionally, we perform a series of ablation studies across different modality settings to validate the contribution of each modality fusion in our approach. We also analyze the impact of contrastive learning on the multi-modal encoder. Finally, we carry out a neighborhood similarity analysis of the health embeddings for living circles across multiple spatial levels.

5.1. Experimental setup

5.1.1. Datasets

Real Estate Data. This dataset is collected from Lianjia,[1] the largest real estate trading platform in China. It contains information on over 8,000 properties in Beijing and over 7,000 in Shanghai. Each property record includes details such as the name of the residential area, street address, latitude and longitude, price, the total number of buildings and units, and corresponding textual descriptions, image links, and quantities. The textual descriptions provide insights into the residential area's environment, covering aspects such as supporting facilities, transportation, surrounding amenities, and target residents. Each residential area also includes a varying number of images that visually depict leisure facilities, environmental greening, and other local features. Additionally, the dataset is supplemented with information from the 2020 permanent population census[2] for both Beijing and Shanghai.

POI Check-in Data. The POI check-in data consists of 753,891 records from Beijing and 549,817 records from Shanghai. The data is sourced from various platforms, including Dianping,[3] Meituan,[4] Baidu Maps,[5] Ctrip,[6] Tujia,[7] and Elong.[8] Dianping is a well-known platform in China for local life services, providing reviews and facilitating transactions. Meituan offers a wide range of services, from dining to entertainment, while Baidu Maps provides travel-related tools such as route planning, navigation, and location queries. Ctrip, Tujia, and Elong are major travel and ticketing platforms where users can leave reviews about their experiences.

The dataset records user check-ins from January 2019 to December 2019, each associated with a unique user ID, venue ID, user comments and ratings, as well as price and source information. Although longitude and latitude information for POIs are not provided, we are able to match POIs with geographical data using the Baidu Maps API, as all platforms in the dataset utilize Baidu Maps for their services. The POIs are categorized into 10 main types, which are selected based on their relevance to chronic disease factors among the elderly. These categories include food, shopping, sports and fitness, tourist attractions, leisure and entertainment, life services, education and training, culture and media, transportation facilities, and stores. The distribution of POIs and reviews across categories is shown in Fig. 3.

---

[1] https://bj.lianjia.com/.

[2] https://www.stats.gov.cn/.

[3] https://www.dianping.com/.

[4] https://www.meituan.com/.

[5] https://map.baidu.com/.

[6] https://hotels.ctrip.com/.

[7] https://m.tujia.com/.
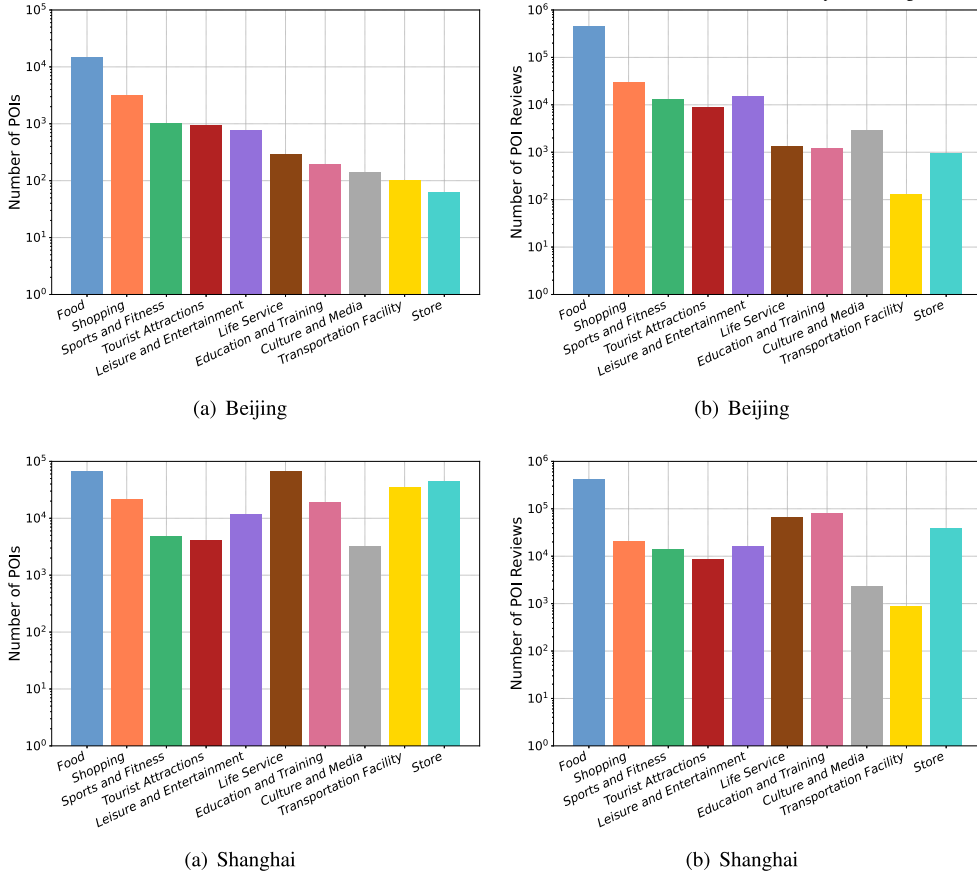
[8] https://www.elong.com/.

Fig. 3. Distribution of POI categories and the number of reviews across the datasets. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Geriatric Disease Data. Due to strict patient privacy protocols and the detailed nature of our research, which focuses on fine-grained residential areas, obtaining comprehensive geriatric disease data presents significant challenges. To address this issue, we use simulated data for tasks related to chronic diseases in the elderly population. We focus on a set of common chronic diseases, including MCI, hypertension, diabetes, and MDD. The prevalence of these diseases is simulated for community residents aged 60 and above in both Beijing and Shanghai. For Beijing, the simulation draws on contrasting morbidity rates between urban core and suburban areas, as reported in medical journals [47,32,61,29]. For Shanghai, the simulation references data from [7,9,18,8]. Additionally, the population of residents aged 60 and above is simulated based on the average household size in each district, as reported in the 2020 population statistics for Beijing and Shanghai, along with the total number of housing units listed on the Lianjia website.

A series of preprocessing steps are implemented to enhance the quality and clarity of the dataset for subsequent tasks. First, to ensure comprehensive multimodal information collection in residential areas, we set a 1000-meter service radius for each area. Next, POI rating samples with a score of 0 within the defined living circle are excluded to ensure compatibility with the text content. Due to the large size of the POI check-in dataset, only a 10% subset of POI review samples from each category is used during the training of the modal encoder. These samples are then further split into testing and validation sets at a 7:3 ratio. The detailed specifications of our datasets are provided in Table 2.

5.1.2. Baselines

To demonstrate the performance of the proposed method, we compare CureGraph against the following algorithms:

- **LINE** [45]. This method optimizes an objective function designed to preserve both local and global network structures using an edge-sampling algorithm. In this study, we apply the LINE technique to multi-modal graphs of living circles and concatenate the embeddings of each modality.
- **Node2Vec** [14]. Node2Vec employs biased random walks to learn latent features that maximize the likelihood of preserving local network structures. Here, we use Node2Vec on multi-modal graphs of living circles and concatenate the embeddings of each modality.
- **GAT** [48]. The Graph Attention Network (GAT) updates the representation of each node by propagating information from its neighbors via an attention mechanism. In this work, we employ a multi-head GAT to learn node representations.

Table 2
Statistics of the experimental datasets for Beijing and Shanghai.

Dataset	Description	
	Beijing	Shanghai
Streets	204 districts divided by administrative streets.	191 districts divided by administrative streets.
Living circles	5087 living circles divided by the 1 km radius of residential areas.	3997 living circles divided by the 1 km radius of residential areas.
Real estate data	5087 residential area texts and around 45 thousand community photos in the living circles.	3997 residential area texts and around 38 thousand community photos in the living circles.
POI check-in data	Over 700 thousand check-in reviews from more than 40 thousand POIs with 10 major categories, along with 143,499 POI locations with 210 middle categories from Baidu Maps.	Over 500 thousand check-in reviews from more than 37 thousand POIs with 10 major categories, along with 275,827 POI locations with 98 middle categories from Baidu Maps.
Geriatric disease data	Prevalence of MCI [47,7]; Prevalence of hypertension [32,9]; Prevalence of diabetes [61,18]; Prevalence of MDD [29,8].	

- **MVGJR** [64]. This unsupervised multi-view joint learning approach is tailored for urban region embedding. It leverages the regional correlations of human mobility and inherent regional properties by employing a cross-view attention mechanism.
- **Urban2Vec** [54]. This is an unsupervised model for neighborhood representation learning. Urban2Vec integrates information from street views and POI data and assumes that neighborhoods with smaller geographic distances are more likely to share common semantics than those farther apart.
- **MuseCL** [63]. This multi-semantic contrastive learning framework utilizes similarities in human mobility patterns and POI distributions within regions to extract semantic features from visual modalities. These visual features are then integrated with textual data from embedded POI texts using a cross-modality attention mechanism and contrastive learning for region representation learning.

5.1.3. Evaluation metrics

To evaluate the performance of the models, we employ three metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to quantify prediction errors, and the coefficient of determination (R^2) to assess the goodness of fit. These metrics are defined as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{Y}_i - Y_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2}, \quad R^2 = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}, \tag{20}$$

where N is the total number of instances, $\hat{Y}_i$ denotes the predicted value, Y_i represents the ground truth, and $\bar{Y}$ is the mean of the observed values. To ensure robust and reliable results, all metrics are calculated using K-Fold cross-validation, with $K = 5$.

5.1.4. Parameters settings

In this section, we detail the experimental settings. For the visual modality encoder, image pairs are randomly selected within the 15-minute living circle for training. The encoder is initialized with parameters pre-trained on ImageNet, with a learning rate of 5×10^{-3} and a temperature hyperparameter τ set to 0.05. For the textual modality encoder, we use BERT as the backbone, pre-trained on a large corpus. The learning rates for the textual augmentation encoder and POI textual modal encoder are set to 1×10^{-5}, with τ set to 1 and 0.005, respectively. The GCN has three layers, a learning rate of 5×10^{-4}, a dropout rate of 0.3, and K set to 20 for the Top-K nearest neighbors. The hyperparameters α, η, and λ are set to 0.2, 0.5, and 0.1, respectively. The batch size is fixed at 32 across all modules. For the Urban2vec model, the embedding size is 200, which is reduced to 128 using PCA, as recommended by the authors. Similarly, the embedding size for LINE, Node2Vec, GAT, MVGJR, and our proposed model and its variants is set to 128.

The implementation is based on PyTorch (version 1.11.0) [40], and all experiments are conducted on an NVIDIA A40 GPU (1.38 GHz, 48 GB GDDR6). We use the Adam optimizer [24] with a weight decay of 3×10^{-3} and train the backbone model for 60 epochs.

5.2. Performance comparison

5.2.1. Overall comparison

This study aims to predict the prevalence of four health conditions—MCI, hypertension, diabetes, and MDD—among elderly individuals aged 60 and above residing in specific neighborhoods. To achieve this, we conduct experiments using datasets from two cities, Beijing and Shanghai, with the learnt urban living circle health embeddings serving as input. The proposed method is evaluated against various baseline models, and the results are summarized in Table 3 and Table 4. In these tables, the best-performing results are highlighted in bold, while the second-best results from baseline models are underlined.

Overall, the results across the four geriatric disease prediction tasks demonstrate the superiority of our proposed CureGraph model. For example, in Beijing and Shanghai, CureGraph achieves improvements of up to 33% and 22% in MAE and RMSE, respectively, while enhancing the R^2 metric by 25% to 145% compared to the best-performing baseline methods.

Among the baselines, MuseCL and Urban2vec demonstrate significant superiority over traditional representation learning methods like Node2Vec and LINE, consistently achieving notable performance boosts. These results suggest that these models are more efficient

Table 3

Prediction results for community elder health outcomes in Beijing across different models.

Model	MCI			Hypertension			Diabetes			MDD		
	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
Node2Vec	23.471	34.277	-0.133	28.534	41.736	-0.106	54.948	82.913	-0.144	17.412	25.007	-0.127
LINE	26.689	37.571	-0.361	31.745	45.241	-0.299	62.904	89.799	-0.315	19.611	27.505	-0.363
GAT	21.600	32.376	-0.010	27.017	39.836	-0.007	51.383	77.767	-0.007	15.996	23.620	-0.005
MVGJR	21.243	32.072	0.003	26.825	39.420	0.002	51.075	77.544	-0.002	15.418	23.610	-0.001
Urban2vec	20.739	31.875	0.044	26.019	39.091	0.040	50.608	77.303	0.026	15.608	23.370	0.033
MuseCL	<u>20.473</u>	<u>31.489</u>	<u>0.052</u>	<u>25.714</u>	<u>38.780</u>	<u>0.045</u>	<u>50.197</u>	<u>76.535</u>	<u>0.029</u>	<u>15.396</u>	<u>23.093</u>	<u>0.039</u>
w/o Top-k sc	20.251	31.237	0.059	25.298	38.380	0.065	49.098	75.593	0.049	15.105	22.827	0.061
Ours	**20.082**	**31.015**	**0.073**	**25.048**	**38.023**	**0.082**	**48.552**	**74.920**	**0.066**	**14.926**	**22.589**	**0.081**

Table 4

Prediction results for community elder health outcomes in Shanghai across different models.

Model	MCI			Hypertension			Diabetes			MDD		
	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
Node2Vec	42.402	57.141	-0.206	153.772	185.692	-0.158	68.702	74.752	-0.141	14.703	18.384	-0.106
LINE	43.231	59.580	-0.291	164.710	195.884	-0.294	73.171	79.897	-0.250	15.457	19.961	-0.374
GAT	41.708	54.283	-0.005	151.209	180.405	-0.009	66.171	73.039	-0.008	13.639	17.866	-0.046
MVGJR	41.946	53.255	0.014	151.065	177.456	0.008	64.275	73.043	0.007	13.221	17.829	0.006
Urban2vec	39.378	52.841	0.046	145.133	170.015	0.055	62.043	71.544	0.047	12.921	17.580	0.033
MuseCL	<u>39.067</u>	<u>52.567</u>	<u>0.056</u>	<u>143.080</u>	<u>167.375</u>	<u>0.068</u>	<u>61.480</u>	<u>70.967</u>	<u>0.058</u>	<u>12.528</u>	<u>17.446</u>	<u>0.058</u>
w/o Top-k sc	38.378	48.608	0.081	139.653	165.866	0.072	60.960	69.422	0.076	11.256	17.298	0.063
Ours	**37.568**	**48.405**	**0.090**	**138.558**	**165.147**	**0.085**	**60.384**	**69.087**	**0.082**	**10.567**	**16.267**	**0.078**

at capturing the complex associations between multimodal spatial data and the health status of the elderly in urban communities. In contrast, LINE's reliance on first-order proximity information and Node2vec's utilization of random walks may be insufficient in capturing the intricate interplay between different modalities and the complex health outcomes of the elderly population. Urban2vec, while effective, depends on distance similarity between images and the fine-tuning of word embeddings for intra-region POI reviews, which poses challenges in aligning image and word semantics. MuseCL, on the other hand, leverages the distributional similarity of POI types to guide the contrastive pairing of images and text. After performing a simple fusion using linear layers, it compares the fused representation with the semantic information of POI types. However, this contrastive fusion approach may inadvertently overemphasize textual semantics, potentially limiting its ability to capture balanced multimodal interactions. Similarly, MVGJR focuses on region correlations derived from region attributes and multi-view learning of POI attributes and check-in data but neglects multimodal semantic interactions. By contrast, our proposed approach captures both the rich semantic dependencies between intersecting modalities and the spatial autocorrelation connections within urban communities. This comprehensive strategy enables our model to outperform others in predicting community health status.

Additionally, we observe that omitting spatial information (e.g., removing Top-k spatial correlation) significantly reduces performance compared to our full approach. This highlights the importance of considering a community's health profile in relation to its surrounding neighborhoods. In summary, our research demonstrates the effectiveness of neighborhood embeddings in health outcome prediction and underscores the necessity of integrating spatial autocorrelation information to improve the accuracy and robustness of these predictions.

5.2.2. Comparison across spatial levels

Our graph construction approach is based on the division of street areas, resulting in a highly interpretable graph representation. This structure enables operations such as averaging, random walks, and pooling, making it versatile for various analytical tasks. In this study, we derive the graph representation of streets by aggregating node representations at the community circle spatial level within the graph. This approach allows us to evaluate whether the proposed CureGraph performs as effectively at the street level as it does at the community level and to explore the similarities and differences between these spatial resolutions. The results in Table 5 and Table 6 demonstrate the superior performance of our proposed CureGraph model at the street level. Despite the coarser granularity, CureGraph consistently achieves state-of-the-art results, with improvements of up to 27% and 32% in MAE and RMSE, respectively, in Beijing and Shanghai, as well as a 5% to 28% increase in the R^2 metric. Notably, for street-level elder health prediction tasks, particularly for conditions such as MDD, CureGraph not only delivers substantial performance gains but also exhibits strong robustness and reliability.

Node2Vec and LINE, in comparison, yield the lowest predictive metrics, highlighting their limited ability to capture complex spatial relationships. The GAT model performs better, dynamically learning relationships between nodes via attention mechanisms, but it falls short in leveraging edge information and managing high-order features. Similarly, MVGJR and Urban2vec fail to fully integrate complementary data from diverse urban morphology and spatial sources. While MuseCL's multimodal contrastive fusion enables it to outperform these models across all metrics, CureGraph surpasses them by employing modality-specific encoders, spatial

Table 5
Elder health prediction results at the street level in Beijing.

Model	MCI			Hypertension			Diabetes			MDD		
	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
Node2Vec	341.615	501.024	0.489	369.277	536.631	0.630	1043.999	1612.160	0.183	227.658	342.144	0.588
LINE	485.110	787.724	-0.264	544.213	883.355	-0.001	983.920	1454.093	0.336	308.203	470.488	0.221
GAT	330.634	490.023	0.508	374.276	561.635	0.430	973.993	1460.159	0.340	215.207	334.477	0.640
MVGJR	327.103	485.723	0.532	373.212	557.356	0.471	977.925	1459.034	0.335	223.207	266.154	0.651
Urban2vec	249.158	361.763	0.693	262.233	389.343	0.783	767.928	1109.036	0.565	151.657	231.478	0.780
MuseCL	<u>244.434</u>	<u>333.503</u>	<u>0.702</u>	<u>254.223</u>	<u>364.427</u>	<u>0.781</u>	<u>574.526</u>	<u>833.201</u>	<u>0.781</u>	<u>142.485</u>	<u>203.176</u>	<u>0.854</u>
w/o Top-k sc	234.147	348.757	0.752	241.200	353.828	0.839	563.449	848.603	0.774	137.272	191.290	0.871
Ours	**229.761**	**346.773**	**0.755**	**239.369**	**343.751**	**0.848**	**560.743**	**822.355**	**0.788**	**130.551**	**180.162**	**0.886**

Table 6
Elder health prediction results at the street level in Shanghai.

Model	MCI			Hypertension			Diabetes			MDD		
	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
Node2Vec	450.866	604.675	0.487	1299.221	1644.190	0.305	1003.658	1685.962	0.297	240.593	390.954	0.528
LINE	549.689	965.301	-0.372	1499.221	1953.103	0.112	1209.286	1750.076	0.171	289.476	495.108	0.321
GAT	440.185	580.970	0.515	1272.914	1572.611	0.435	801.433	1629.305	0.342	236.891	340.864	0.588
MVGJR	436.081	572.376	0.539	1153.373	1531.367	0.492	790.271	1607.493	0.369	229.492	335.456	0.591
Urban2vec	358.976	451.309	0.685	1118.217	1438.046	0.605	650.163	1433.858	0.503	156.802	320.854	0.719
MuseCL	<u>349.883</u>	<u>430.877</u>	<u>0.694</u>	<u>1036.118</u>	<u>1259.877</u>	<u>0.682</u>	<u>642.501</u>	<u>1338.832</u>	<u>0.602</u>	<u>146.594</u>	<u>302.387</u>	<u>0.726</u>
w/o Top-k sc	338.083	415.842	0.738	966.816	1043.490	0.781	579.419	1112.819	0.689	138.947	290.874	0.745
Ours	**324.790**	**410.314**	**0.742**	**869.192**	**1004.493**	**0.823**	**572.512**	**984.042**	**0.765**	**135.495**	**384.592**	**0.762**

synthesis techniques, and effectively capturing both spatial autocorrelation within neighborhoods and spatial heterogeneity across them.

Furthermore, the R^2 metric for elder health prediction highlights CureGraph's superior fitting capability at the street level compared to the community living circle level. This can be attributed to our graph construction methodology, which accounts for spatial dependencies among various community patterns within neighborhoods. Interestingly, improvements in MAE and RMSE metrics at the street level are more pronounced than at the community level. This suggests that CureGraph exhibits heightened sensitivity to spatial autocorrelation at finer spatial resolutions. However, this sensitivity may also amplify the effects of outliers or anomalous data points, influencing these metrics more significantly at detailed spatial levels.

5.3. Modality ablation study

In this subsection, we conduct a modality ablation study to evaluate the contributions of individual modalities and their combinations within our proposed multi-modal model, CureGraph. Tables 7 and 8 present CureGraph's performance under various multi-modal configurations for datasets from Beijing and Shanghai, respectively. Additionally, Fig. 4 provides a visual representation of the R^2 values for community elder health predictions in these two cities.

More specifically, the study examines prediction performance across different settings, including single-modality variants (**T-Only**, **V-Only**, **P-Only**) and combined-modality variants (**w/o T, w/o V, w/o P**). Here, **T** represents the textual modality, **V** denotes the visual modality, and **P** refers to the POI review modality. The results reveal significant variations in predictive performance across these modalities. Notably, the visual modality (V-Only) demonstrates strong predictive capabilities, particularly for MCI and hypertension in Beijing. Furthermore, combining the textual and visual modalities (w/o P) improves predictions for diabetes and MDD, achieving an average improvement of 112.9% in R^2 across datasets from Beijing and Shanghai. Conversely, the POI review modality (P-Only) exhibits limited predictive power and does not significantly enhance outcomes when included.

Despite variability in the performance of individual modalities, combining all three modalities consistently yields the highest predictive accuracy (represented by the blue bar for our method in Fig. 4). This improvement can be attributed to the alignment of POI text embeddings with residential textual and visual modalities during multimodal fusion, resulting in better integration within the vector space. These findings emphasize the critical role of multimodal fusion in enhancing predictive accuracy for elderly health conditions.

Additionally, we observe that straightforward combinations of multimodal embeddings, as employed by baseline models like LINE and Node2Vec, perform worse than single-modality approaches. As shown in Table 3, these baseline models fail to match the performance of single-modality methods across all four chronic disease prediction tasks. This highlights the inadequacy of simplistic multimodal embedding techniques in capturing the complex interactions required for accurate elderly health predictions.

Table 7
Ablation study results for community-level elder health prediction in Beijing.

Model	MCI			Hypertension			Diabetes			MDD		
	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
T-Only	21.200	31.990	0.014	26.491	39.398	0.015	50.655	77.287	0.006	15.718	23.406	0.013
V-Only	20.760	31.904	0.019	26.017	39.277	0.021	50.720	77.342	0.004	15.597	23.377	0.015
P-Only	21.486	32.250	-0.003	26.906	39.742	-0.003	51.282	77.651	-0.004	15.970	23.594	-0.002
w/o T	20.982	32.198	0.001	26.255	39.521	0.008	50.803	77.416	0.002	15.675	23.424	0.012
w/o V	21.418	32.271	-0.004	26.768	39.710	-0.001	50.893	77.483	0.001	15.866	23.548	0.001
w/o P	21.232	32.076	0.008	26.573	39.433	0.012	50.589	76.947	0.015	15.745	23.364	0.017
Ours	**20.082**	**31.015**	**0.073**	**25.048**	**38.023**	**0.082**	**48.552**	**74.920**	**0.066**	**14.926**	**22.589**	**0.081**

Table 8
Ablation study results for community-level elder health prediction in Shanghai.

Model	MCI			Hypertension			Diabetes			MDD		
	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
T-Only	41.786	53.229	0.016	149.803	175.873	0.018	64.052	72.893	0.011	13.154	17.774	0.012
V-Only	41.464	53.067	0.022	150.182	176.348	0.015	64.219	73.006	0.008	13.143	17.764	0.013
P-Only	41.796	53.904	0.002	151.209	180.405	-0.009	64.486	73.041	0.002	13.293	17.835	-0.003
w/o T	41.871	53.580	0.008	151.065	177.456	0.008	63.773	72.706	0.016	13.154	17.774	0.012
w/o V	41.846	53.688	0.006	151.090	177.976	0.006	63.996	72.856	0.012	13.245	17.831	0.003
w/o P	41.625	53.203	0.018	149.550	175.556	0.020	63.147	72.362	0.018	13.088	17.718	0.018
Ours	**37.568**	**48.405**	**0.090**	**138.558**	**165.147**	**0.085**	**60.384**	**69.087**	**0.082**	**10.567**	**16.267**	**0.078**
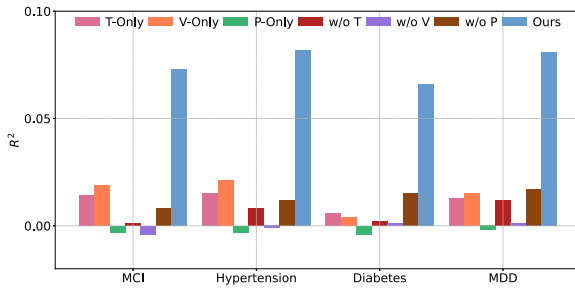
(a) Beijing
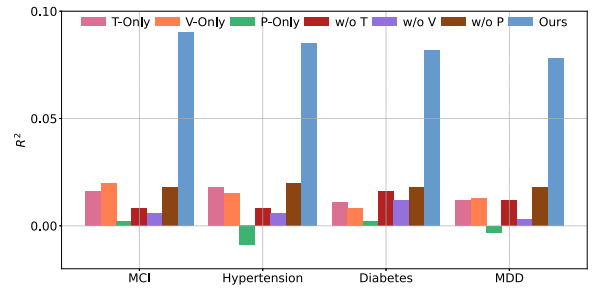
(b) Shanghai

Fig. 4. Ablation study results for R^2 in community elder health prediction.

5.4. Analysis of multi-modal encoder

To illustrate the effects of the contrastive learning-based modality encoder and uncover semantic meanings, we visualize the multi-modal embeddings of urban living circles using both pre-trained model methods and our contrastive learning encoder methods. Fig. 5(a) and Fig. 5(c) employ pre-trained models (DenseNet [19] for images and BERT for texts, respectively), while Fig. 5(b) utilizes a self-supervised contrastive learning approach for text encoding. Principal Component Analysis (PCA) is applied to project the features into a two-dimensional feature space, as shown in Fig. 5.

The color of each dot corresponds to the population values of geriatric diseases within urban communities. The comparison reveals that the contrastive learning approach for encoding visual and POI text modalities increases the separation in the feature space of living circles, effectively amplifying differences among modalities. Regarding the text modality of residential areas in Fig. 5(e), we observe that the first principal component is associated with the number of bus lines, where living circles with multiple bus lines exhibit lower values along this dimension. Furthermore, the text embeddings form two distinct clusters: one primarily associated with residential areas and the other with higher values on both the first and second components, suggesting a strong relationship with transportation and surrounding facilities. Notably, the latter cluster lacks textual descriptions of internal environments, indicating that our modality encoder successfully reduces the distance between semantically similar modalities while increasing the distance between dissimilar ones, particularly when incorporating residential area image semantics through cross-modal contrast learning.

Furthermore, we compute the pairwise cosine similarities of text representation vectors from POI reviews and present the results as heatmaps in Fig. 6. Specifically, Fig. 6(a) illustrates the results obtained using our proposed POI Textual Modality Encoder, while Fig. 6(b) shows those derived from BERT-based representations of the same texts. For this analysis, we randomly select 100 POI review samples, each comprising 20 reviews sorted by rating (from 1 to 5) in ascending order. Fig. 6(b) demonstrates that the pre-trained
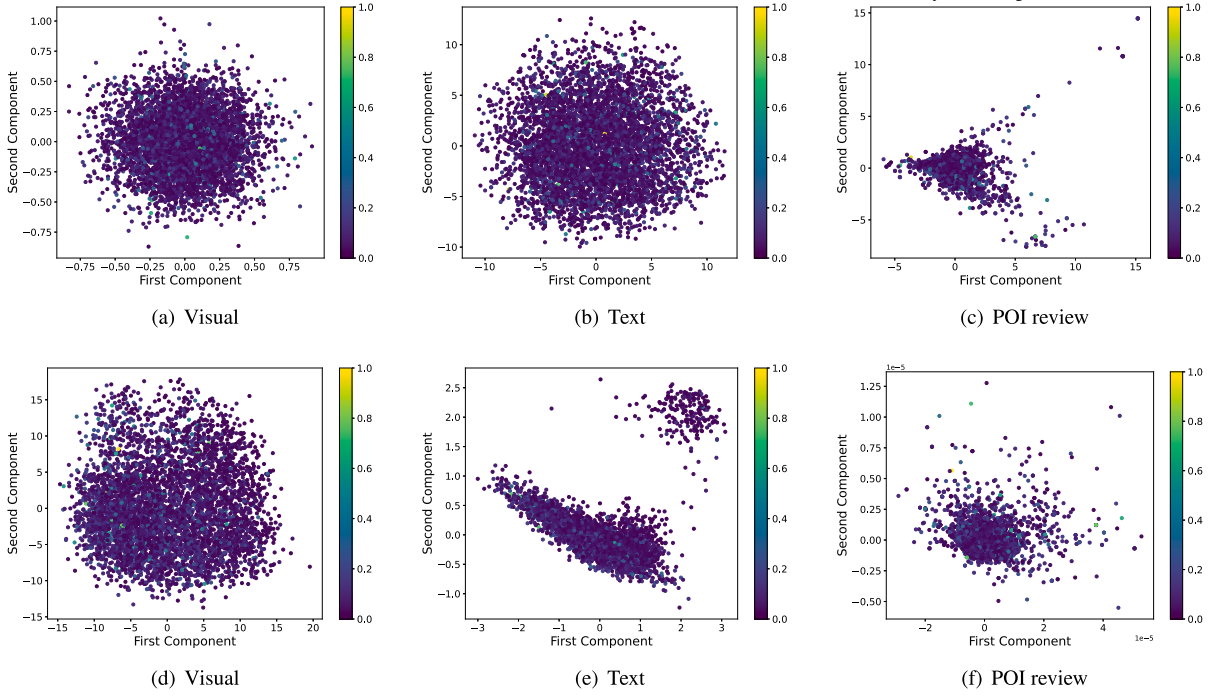
(a) Visual   (b) Text   (c) POI review

(d) Visual   (e) Text   (f) POI review

Fig. 5. PCA-based visualization of urban multi-modal embeddings.

(a) Similarity of reviews based on POI Textual Modality Encoder
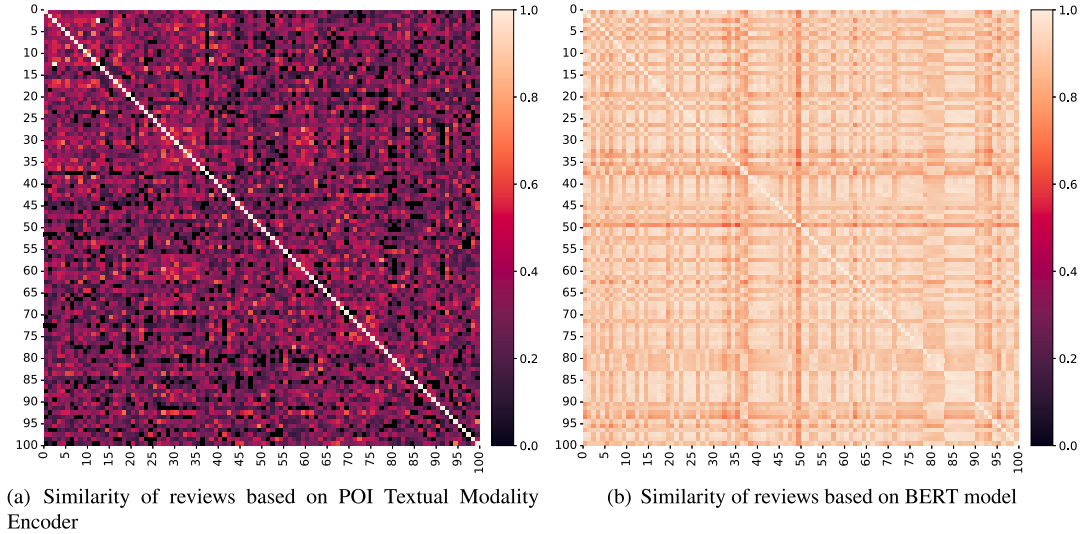
(b) Similarity of reviews based on BERT model

Fig. 6. Heatmaps of POI review representation similarities.

language model BERT produces uniformly high textual similarity, with nearly all sentence pairs achieving scores between 0.7 and 1.0. In contrast, Fig. 6(a) reveals that the supervised contrastive learning encoder generates more pronounced disparities between sentence pairs, effectively contracting the similarity of positive samples while expanding the divergence for negative samples. Notably, reviews with lower ratings near the diagonal exhibit distinctly high similarity, underscoring the encoder's ability to capture subtle distinctions between reviews.

5.5. Analysis of modality impact

To better understand the influence of multimodal embeddings on the prediction outcomes for elderly diseases, we employ Shapley Additive Explanations (SHAP) values [31] to evaluate the importance of various features. SHAP values quantify the contribution of individual features to specific predictions, providing deeper insights into how different features derived from multiple modalities affect
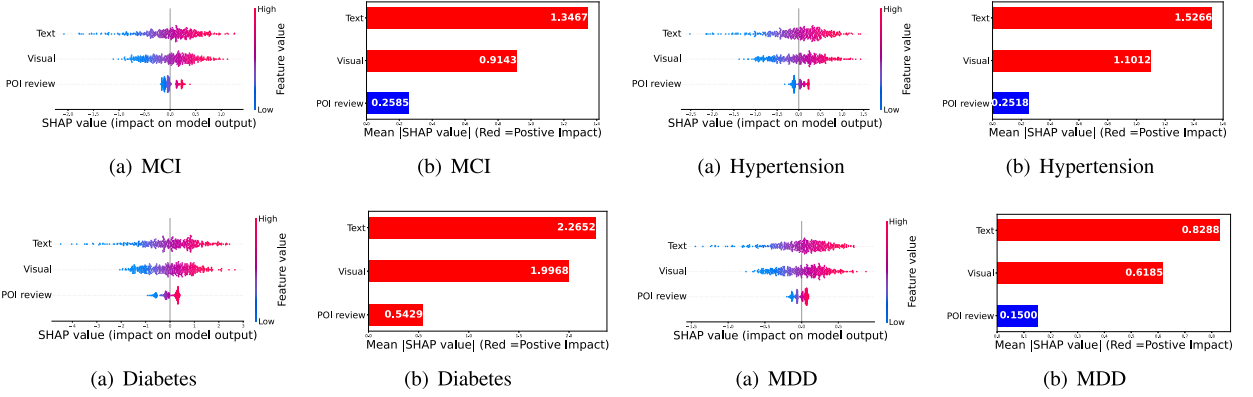
Fig. 7. SHAP value plots for prediction analysis. (a) A summary of SHAP values for modal embeddings. (b) Average absolute SHAP values, with color indicating positive (red) or negative (blue) impacts on the prediction outcomes.

CureGraph's outputs. To ensure computational efficiency, we randomly select embedding data from 500 communities in Beijing to assess their roles in the regression model's predictions.

In Fig. 7, panel (a) illustrates the SHAP values for different modalities in predicting each elderly disease. Modal features are represented as horizontal lines, with colors indicating the magnitude of feature values: red points denote higher feature values, while blue points represent lower values. Panel (b) shows the average absolute SHAP value for each feature, along with the correlation between feature values and their SHAP values. Features depicted by red bars have positive correlation coefficients, meaning that higher feature values are associated with an increased predicted number of elderly individuals with the disease. Conversely, features shown with blue bars exhibit negative correlations, where higher feature values correspond to a decreased predicted number of elderly individuals with the disease.

Among the four diseases analyzed, the text modality learned by CureGraph has the most significant impact on prediction outcomes, followed by the visual modality, with POI review data having the least influence. Interestingly, compared to the other two modalities, higher POI review feature values are generally positively correlated with the number of elderly individuals suffering from MCI, hypertension, and MDD in the community, but negatively correlated with the number of elderly individuals suffering from diabetes.

5.6. Analysis of health disparities

5.6.1. Health differences at the community level

For a given living circle within a city, we compute the cosine similarity between its embedding vector and those of other neighborhoods in the city. Based on these similarity metrics, we rank the neighborhoods. In Fig. 8, we present two examples of neighborhood comparisons with cosine similarity scores of 0.93 and -0.26, respectively.

As shown in Fig. 8, urban living circle A, presented in the first row, and its most similar neighborhood (second row) exhibit striking resemblances in physical appearance and living environment. Both areas are characterized by convenient transportation, low floor area ratios, abundant green spaces, and high housing prices. Additionally, they share similarities in the number of residential households and the distribution of the elderly population. In contrast, the community least similar to urban living circle A (third row) displays distinct features, such as lower housing values, limited green spaces, and significant challenges related to aging. This community faces a substantial elderly population combined with a low coverage rate of relevant service facilities.

5.6.2. Health differences at the street level

To evaluate health disparities at the street level, we apply K-means clustering on embeddings generated by our model for Beijing. Fig. 9(a) displays the clustering results with $k = 3$, determined by the elbow method. Fig. 9(b) depicts the distribution of the population aged 60 and above in Beijing. As shown in the figure, a notable cluster 0 (green) emerges around streets with high facility coverage outside the city's second ring road and in certain suburban areas. These regions are characterized by high population mobility, frequent activities, and dense urban development. Conversely, cluster 1 (purple) encompasses sparsely populated areas, indicating low concentrations of elderly individuals. Within the second ring road, the Dongcheng and Xicheng districts exhibit a distinct clustering pattern, underscoring a significant aging demographic challenge. These central areas are outlined by a red-bordered boundary. Streets in the orange cluster predominantly represent extensive, underdeveloped township administrative regions, discernible from their delineated boundaries. Fig. 9(b) further highlights that these regions contain the highest concentrations of elderly individuals.

Our analysis indicates comparable health representations and elderly population distributions between suburban and central urban areas. However, Beijing municipal census data reveals a significant aging challenge in the city center. In districts such as Dongcheng and Xicheng, the proportion of the population aged 60 and above reaches approximately 26%. These central communities urgently require elderly-friendly facilities and supportive welfare policies to address the needs of their aging residents.
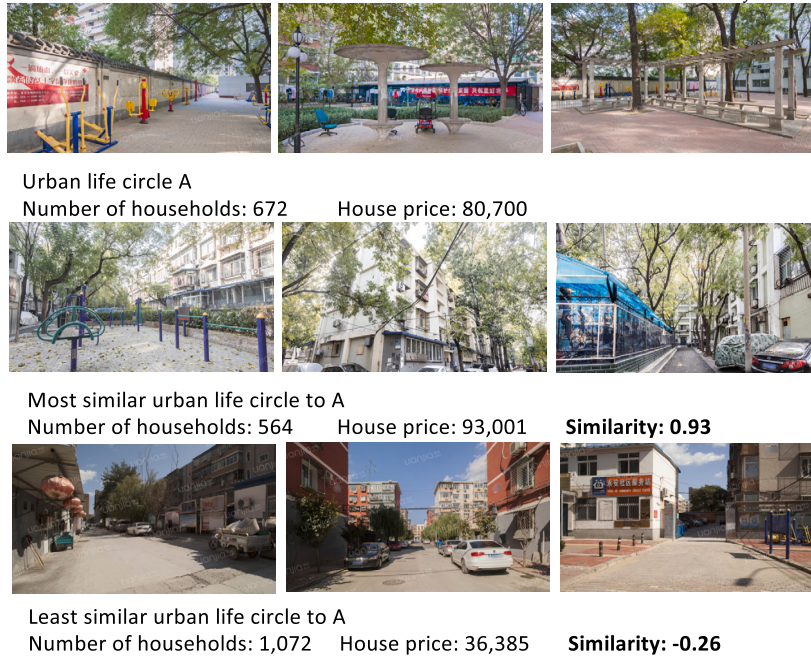
Urban life circle A
Number of households: 672　　　House price: 80,700

Most similar urban life circle to A
Number of households: 564　　　House price: 93,001　　　**Similarity: 0.93**

Least similar urban life circle to A
Number of households: 1,072　　House price: 36,385　　　**Similarity: -0.26**

Fig. 8. Community health differences in Beijing based on living circle embeddings.

(a) K-means clustering of street embeddings in Beijing(K=3)　　　(b) Distribution of the population aged 60 and above in Beijing
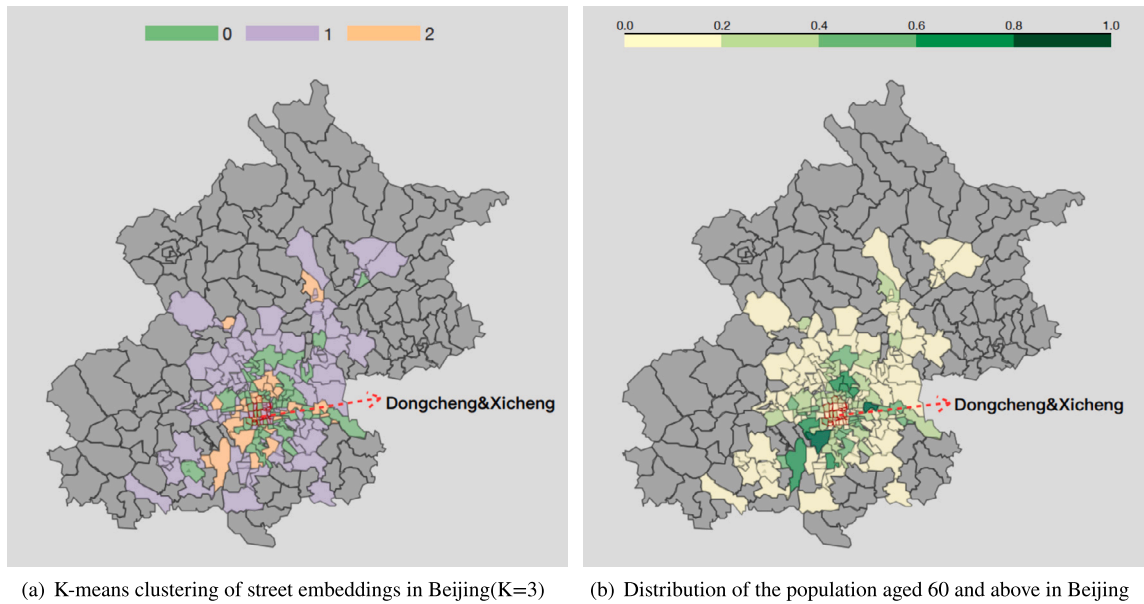
Fig. 9. Cluster and choropleth maps of street embeddings and elderly population in Beijing.

5.6.3. Socioeconomic bias

To investigate potential socioeconomic biases, we conduct a correlation analysis examining the relationship between predicted elderly disease prevalence and socioeconomic factors. Using the Pearson correlation coefficient (PCC) [5] and p-value, we assess the association between predicted disease prevalence and housing prices. The results indicate a significant negative correlation (PCC = $-0.17, p < 0.05$), suggesting that communities with higher housing prices tend to have fewer predicted elderly individuals with diseases. This finding may reflect biases in the data collection process. Communities with higher housing prices often feature higher-quality images and more detailed textual descriptions, which can enhance the model's predictive performance in these areas. Conversely, economically disadvantaged communities may suffer from lower data quality or insufficient details, resulting in relatively less accurate predictions. These disparities highlight the importance of addressing data quality issues to ensure equitable model performance across different socioeconomic contexts.

6. Conclusion

In this work, we introduce **CureGraph**, an innovative unsupervised graph-based multi-modal representation learning framework that leverages photos, textual reviews of residential areas, and surrounding POI data to generate urban neighborhood health embeddings. Our framework constructs a graph to capture cross-modal dependencies and spatial autocorrelation among neighborhoods, enabling the effective fusion of diverse data modalities. This approach allows for fine-grained predictions of community elder health statuses and uncovers urban health disparities across multiple spatial levels. We validate CureGraph through experiments on real-world datasets and extensive ablation studies under various modality settings. The findings offer valuable insights for urban planners and policymakers, fostering healthier urban communities and supporting spatial planning for healthy aging.

Despite its strengths, our approach has certain limitations. First, it relies predominantly on static images and textual data, limiting its ability to capture the dynamic nature of urban environments and their continuous evolution. Second, while our framework accounts for spatial autocorrelation among urban regions, it does not fully incorporate road network data. This is a significant limitation, as road networks play a crucial role in revealing intricate connections and interaction patterns within cities, profoundly influencing mobility, lifestyle behaviors, and community interactions.

To address these limitations, future work will focus on incorporating dynamic data sources to better capture urban vibrancy and change. We also plan to integrate road network data into the framework to gain a deeper understanding of urban spatial dynamics and complexities. To enhance the inclusiveness and precision of CureGraph, we aim to design novel algorithms that can adapt to evolving urban contexts. Furthermore, with a strong emphasis on privacy protection, we will explore anonymized or pseudo-anonymized data collection methods and collaborate with healthcare institutions and community service organizations. Accessing mobility and health data specific to the elderly population will enable the provision of more targeted and effective services with greater precision. These advancements are expected to significantly enhance insights into urban planning and governance, promoting environments that support healthy aging and fostering improved community welfare.

CRediT authorship contribution statement

Jinlin Li: Writing – original draft, Methodology, Data curation. **Xiao Zhou:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgement

Data availability

Data will be made available on request.

References

[1] Carol S. Aneshensel, Frederick Harig, Richard G. Wight, Aging, neighborhoods, and the built environment, in: Handbook of Aging and the Social Sciences, Elsevier, 2016, pp. 315–335.
[2] Luc Anselin, What is special about spatial data? Alternative perspectives on spatial data analysis (89-4), 1989.
[3] Marica Cassarino, Annalisa Setti, Environment as 'brain training': a review of geographical and physical environmental influences on cognitive ageing, Ageing Res. Rev. 23 (2015) 167–182.
[4] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, Yaliang Li, Simple and deep graph convolutional networks, in: International Conference on Machine Learning, PMLR, 2020, pp. 1725–1735.
[5] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, Israel Cohen, Pearson correlation coefficient, in: Noise Reduction in Speech Processing, 2009, pp. 1–4.
[6] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Pre-training with whole word masking for Chinese bert, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 3504–3514, https://doi.org/10.1109/TASLP.2021.3124365.
[7] Ding Ding, Qianhua Zhao, Qihao Guo, Haijiao Meng, Bei Wang, Jianfeng Luo, James A. Mortimer, Amy R. Borenstein, Zhen Hong, Prevalence of mild cognitive impairment in an urban community in China: a cross-sectional analysis of the Shanghai aging study, Alzheimer's Dement. 11 (3) (2015) 300–309.
[8] Shuhui Dong, Xirong Sun, Tao Wang, Shifu Xiao, Investigation and analysis of depressive symptoms in elderly of community, J. Intern. Med. Concepts Pract. 13 (03) (2018) 170–174.
[9] Weizhen Dong, Jin Wan, Yanjun Xu, Chun Chen, Ge Bai, Lyuying Fang, Anjiang Sun, Yinghua Yang, Ying Wang, Determinants of self-rated health among Shanghai elders: a cross-sectional study, BMC Public Health 17 (2017) 1–12.
[10] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, Xiaolin Li, Efficient region embedding with multi-view spatial networks: a perspective of locality-constrained spatial autocorrelations, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 906–913.

[11] Panagis Galiatsatos, Cynthia Kineza, Seungyoun Hwang, Juliana Pietri, Emily Brigham, Nirupama Putcha, Cynthia S. Rand, Meredith McCormack, Nadia N. Hansel, Neighbourhood characteristics and health outcomes: evaluating the association between socioeconomic status, tobacco store density and health outcomes in Baltimore city, Tob. Control 27 (e1) (2018) e19–e24.

[12] Tianyu Gao, Xingcheng Yao, Danqi Chen, Simcse: simple contrastive learning of sentence embeddings, arXiv preprint arXiv:2104.08821, 2021.

[13] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, George E. Dahl, Neural message passing for quantum chemistry, in: International Conference on Machine Learning, PMLR, 2017, pp. 1263–1272.

[14] Aditya Grover, Jure Leskovec, node2vec: scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 855–864.

[15] Yingqi Guo, Yuqi Liu, Shiyu Lu, On Fung Chan, Cheryl Hiu Kwan Chui, Terry Yat Sang Lum, Objective and perceived built environment, sense of community, and mental wellbeing in older adults in Hong Kong: a multilevel structural equation study, Landsc. Urban Plan. 209 (2021) 104058.

[16] Apinan Hasthanasombat, Cecilia Mascolo, Understanding the effects of the neighbourhood built environment on public health with open data, in: The World Wide Web Conference, 2019, pp. 648–658.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[18] Xing Hu, Lingbing Meng, Zhimin Wei, Hongxuan Xu, Jianyi Li, Yingying Li, Na Jia, Hui Li, Xin Qi, Xuezhai Zeng, et al., Prevalence and potential risk factors of self-reported diabetes among elderly people in China: a national cross-sectional study of 224,142 adults, Front. Public Health 10 (2022).

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[20] Le Huang, Fan Xia, Hui Chen, Bowen Hu, Xiao Zhou, Chunxiao Li, Yaohui Jin, Yanyan Xu, Reconstructing human activities via coupling mobile phone data with location-based social networks, Travel Behav. Soc. 33 (2023) 100606.

[21] Tianyuan Huang, Zhecheng Wang, Hao Sheng, Andrew Y. Ng, Ram Rajagopal, Learning neighborhood representation from multi-modal multi-graph: image, text, mobility graph and beyond, arXiv preprint arXiv:2105.02489, 2021.

[22] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, Fillia Makedon, A survey on contrastive self-supervised learning, Technologies 9 (1) (2020) 2.

[23] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, Stefano Ermon, Tile2vec: unsupervised representation learning for spatially distributed data, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3967–3974.

[24] Diederik P. Kingma, Jimmy Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[25] Li Tong, Shiduo Xin, Yanxin Xi, Sasu Tarkoma, Pan Hui, Yong Li, Predicting multi-level socioeconomic indicators from structural urban imagery, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 3282–3291.

[26] Xin Li, Chao Ma, Junying Zhang, Ying Liang, Yaojing Chen, Kewei Chen, Jun Wang, Zhanjun Zhang, Yongyan Wang, Beijing Ageing Brain Rejuvenation Initiative, Prevalence of and potential risk factors for mild cognitive impairment in community-dwelling residents of Beijing, J. Am. Geriatr. Soc. 61 (12) (2013) 2111–2119.

[27] Yi Li, Weiming Huang, Gao Cong, Hao Wang, Zheng Wang, Urban region representation learning with openstreetmap building footprints, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 1363–1373.

[28] Hongxiao Liu, Feng Li, Juanyong Li, Yuyang Zhang, The relationships between urban parks, residents' physical activity, and mental health benefits: a case study from Beijing, China, J. Environ. Manag. 190 (2017) 223–230.

[29] Jing Liu, Fang Yan, Xin Ma, Hong-Li Guo, Yi-Lang Tang, Jeffrey J. Rakofsky, Xiao-Mei Wu, Xiao-Qiang Li, Hong Zhu, Xiao-Bing Guo, et al., Prevalence of major depressive disorder and socio-demographic correlates: results of a representative household epidemiological survey in Beijing, China, J. Affect. Disord. 179 (2015) 74–81.

[30] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, Yong Li, Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 4150–4160.

[31] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, Su-In Lee, From local explanations to global understanding with explainable ai for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67.

[32] Lina Ma, Xiaoling Zhao, Zhe Tang, Yun Li, Fei Sun, Lijun Diao, Gaizhen Ge, Ming Feng, Jieyu Wang, Epidemiological characteristics of hypertension in the elderly in Beijing, China, PLoS ONE 10 (8) (2015) e0135480.

[33] Wenjun Ma, Ning Wang, Yuxi Li, Daniel Jian Sun, 15-min pedestrian distance life circle and sustainable community governance in Chinese metropolitan cities: a diagnosis, Humanit. Soc. Sci. Commun. 10 (1) (2023) 1–14.

[34] Adyasha Maharana, Elaine Okanyene Nsoesie, Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity, JAMA Netw. Open 1 (4) (2018) e181535.

[35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013).

[36] Harvey J. Miller, Tobler's first law and spatial analysis, Ann. Assoc. Am. Geogr. 94 (2) (2004) 284–289.

[37] Carlos Moreno, Zaheer Allam, Didier Chabaud, Catherine Gall, Florent Pratlong, Introducing the "15-minute city": sustainability, resilience and place identity in future post-pandemic cities, Smart Cities 4 (1) (2021) 93–111.

[38] Aaron van den Oord, Yazhe Li, Oriol Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748, 2018.

[39] World Health Organization, et al., Decade of healthy ageing: baseline report, 2020.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., Pytorch: an imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).

[41] Ewa Rudnicka, Paulina Napierała, Agnieszka Podfigurna, Błażej Męczekalski, Roman Smolarczyk, Monika Grymowicz, The world health organization (who) approach to healthy ageing, Maturitas 139 (2020) 6–11.

[42] Konstantinos Skianis, Fragkiskos Malliaros, Michalis Vazirgiannis, Fusing document, collection and label graph-based representations with word embeddings for text classification, in: Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12), 2018, pp. 49–58.

[43] Jian Song, Peng Du, Weizhuo Yi, Jing Wei, Jianlong Fang, Rubing Pan, Feng Zhao, Yi Zhang, Zhiwei Xu, Qinghua Sun, et al., Using an exposome-wide approach to explore the impact of urban environments on blood pressure among adults in Beijing–Tianjin–Hebei and surrounding areas of China, Environ. Sci. Technol. 56 (12) (2022) 8395–8405.

[44] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[45] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, Qiaozhu Mei, Line: large-scale information network embedding, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 1067–1077.

[46] Jie Tang, Nanqian Chen, Hailun Liang, Xu Gao, The effect of built environment on physical health and mental health of adults: a nationwide cross-sectional study in China, Int. J. Environ. Res. Public Health 19 (11) (2022) 6492.

[47] Zhe Tang, Xinqing Zhang, W.U. Xiaoguang, Prevalence of the mild cognitive impairment among elderly in Beijing, Chin. Ment. Health J. (1991).

[48] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, Graph attention networks, Stat 1050 (20) (2017).

[49] Hongjian Wang, Zhenhui Li, Region representation learning via mobility flow, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 237–246.

[50] Ruoyu Wang, Yi Lu, Jinbao Zhang, Penghua Liu, Yao Yao, Ye Liu, The relationship between visual enclosure for neighbourhood street walkability and elders' mental health in China: using street view images, J. Transp. Health 13 (2019) 90–102.

[51] Tongzhou Wang, Phillip Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere, in: International Conference on Machine Learning, PMLR, 2020, pp. 9929–9939.

[52] Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Lichao Mou, Xiao Xiang Zhu, Self-supervised learning in remote sensing: a review, arXiv preprint arXiv:2206.13188, 2022.

[53] Yingzi Wang, Xiao Zhou, Anastasios Noulas, Cecilia Mascolo, Xing Xie, Enhong Chen, Predicting the spatio-temporal evolution of chronic diseases in population with human mobility data, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 3578–3584.

[54] Zhecheng Wang, Haoyuan Li, Ram Rajagopal, Urban2vec: incorporating street view imagery and pois for multi-modal urban neighborhood embedding, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 1013–1020.

[55] Janine L. Wiles, Ruth E.S. Allen, Anthea J. Palmer, Karen J. Hayman, Sally Keeling, Ngaire Kerse, Older people and their social spaces: a study of well-being and attachment to place in aotearoa New Zealand, Soc. Sci. Med. 68 (4) (2009) 664–671.

[56] Janine L. Wiles, Annette Leibing, Nancy Guberman, Jeanne Reeve, Ruth ES Allen, The meaning of "aging in place" to older people, Gerontologist 52 (3) (2012) 357–366.

[57] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, Hao Ma, Clear: contrastive learning for sentence representation, arXiv preprint arXiv:2012.15466, 2020.

[58] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, Pan Hui, Beyond the first law of geography: learning representations of satellite imagery by leveraging point-of-interests, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3308–3316.

[59] Zhuo Xu, Xiao Zhou, Cgap: urban region representation learning with coarsened graph attention pooling, in: Proceedings of the Thirty Third International Joint Conference on Artificial Intelligence, IJCAI-24, 2024, pp. 7518–7526.

[60] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, Weiran Xu, Consert: a contrastive framework for self-supervised sentence representation transfer, arXiv preprint arXiv:2105.11741, 2021.

[61] Ze Yang, Hong Zheng, Zhifu Tong, Prevalent status of diabetes and igt in senile population of Beijing in 1997, Chin. J. Geriatr. (2000).

[62] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, Hui Xiong, Representing urban functions through zone embedding with human mobility patterns, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 2018.

[63] Xixian Yong, Xiao Zhou, Musecl: predicting urban socioeconomic indicators via multi-semantic contrastive learning, in: Proceedings of the Thirty Third International Joint Conference on Artificial Intelligence, IJCAI-24, 2024, pp. 7536–7544.

[64] Mingyang Zhang, Tong Li, Yong Li, Pan Hui, Multi-view joint graph representation learning for urban region embedding, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 4431–4437.

[65] Xiao Zhou, Anastasios Noulas, Cecilia Mascolo, Zhongxiang Zhao, Discovering latent patterns of urban cultural interactions in wechat for modern city planning, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1069–1078.

[66] Xiao Zhou, Xiaohu Zhang, Paolo Santi, Carlo Ratti, Phase-wise evaluation and optimization of non-pharmaceutical interventions to contain the covid-19 pandemic in the US, Front. Public Health 11 (2023) 1198973.