Senseable City Lab :.:: Massachusetts Institute of Technology

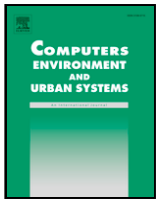
This paper might be a pre-copy-editing or a post-print author-produced .pdf of an article accepted for publication. For the definitive publisher-authenticated version, please refer directly to publishing house's archive system

Computers, Environment and Urban Systems xxx (xxxx) 102043

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



Comprehensive urban space representation with varying numbers of street-level images

Yingjing Huang [a, b], Fan Zhang [a, *], Yong Gao [a], Wei Tu [c], Fabio Duarte [b], Carlo Ratti [b], Diansheng Guo [d], Yu Liu [a]

[a] Institute of Remote Sensing and Geographical Information System, School of Earth and Space Sciences, Peking University, Beijing, China
[b] Senseable City Lab, Massachusetts Institute of Technology, United States
[c] Guangdong Key Laboratory of Urban Informatics, Research Institute for Smart Cities, and Shenzhen Key Laboratory of Spatial Smart Sensing and Services, and Department of Urban Informatics, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China
[d] Tencent Corporation, Beijing, China

UNCORRECTED PROOF

ARTICLE INFO

Keywords:
Street-level imagery
Urban space representation
Multimodal data fusion
Deep learning
Urban village recognition

ABSTRACT

Street-level imagery has emerged as a valuable tool for observing large-scale urban spaces with unprecedented detail. However, previous studies have been limited to analyzing individual street-level images. This approach falls short in representing the characteristics of a spatial unit, such as a street or grid, which may contain varying numbers of street-level images ranging from several to hundreds. As a result, a more comprehensive and representative approach is required to capture the complexity and diversity of urban environments at different spatial scales. To address this issue, this study proposes a deep learning-based module called Vision-LSTM, which can effectively obtain vector representation from varying numbers of street-level images in spatial units. The effectiveness of the module is validated through experiments to recognize urban villages, achieving reliable recognition results (overall accuracy: 91.6%) through multimodal learning that combines street-level imagery with remote sensing imagery and social sensing data. Compared to existing image fusion methods, Vision-LSTM demonstrates significant effectiveness in capturing associations between street-level images. The proposed module can provide a more comprehensive understanding of urban spaces, enhancing the research value of street-level imagery and facilitating multimodal learning-based urban research. Our models are available at https://github.com/yingjinghuang/Vision-LSTM.

1. Introduction

Urban spaces have become increasingly complex and challenging to study due to the rapid growth of cities. Recent years have seen a significant transformation in urban research with the emergence of a new data source, street-level imagery, which has been widely adopted as a novel tool for observing large-scale urban environments in unprecedented detail (Biljecki & Ito, 2021; Duarte & Ratti, 2021; Ibrahim, Haworth, & Cheng, 2020). Compared to remote sensing imagery, street-level imagery offers a human-like perspective of urban spaces, providing a unique view of the environment that cannot be achieved through a nadir view (Chen et al., 2022). With recent advancements in deep learning and computer vision techniques, researchers can now automatically and efficiently extract high-level semantic information from street-level imagery. These latent features not only provide visual cues that shape human experiences of urban spaces, but also offer valuable insights into the socioeconomic status and human dynamics within cities (Fan, Zhang, Loo, & Ratti, 2023; Khosla, An An, Lim, & Torralba, 2014; Zhang, Wu, Zhu, & Liu, 2019). The growing importance of street-level imagery in urban research and its ability to provide fine-grained details of urban spaces has propelled it to the forefront of research in urban planning and policymaking.

Previous studies have predominantly focused on analyzing individual street-level images or sample points (Seiferling, Naik, Ratti, & Proulx, 2017; Zhang, Zhou, Ratti, & Liu, 2019). For instance, the study conducted by Zhang et al. (2018) examined human perceptions based on individual street-level images, while Li et al. (2015) employed street-level panorama imagery to compute the green view index at sampling points. Such approach can only capture the local characteristics of the urban scenes but falls short in representing the characteristics of a spatial unit, such as a street, grid, or block (Feng et al., 2021; Huang, Yang, Li, & Wen, 2021). However, urban landscapes and urban func-

* Corresponding author.
E-mail address: fanzhanggis@pku.edu.cn (F. Zhang).

https://doi.org/10.1016/j.compenvurbsys.2023.102043
Received 4 May 2023; Received in revised form 21 September 2023; Accepted 23 September 2023
0198-9715/© 20XX

Note: Low-resolution images were used to create this PDF. The original images will be used in the final composition.

tions are organized based on spatial units at varying spatial scales. By analyzing street-level imagery in spatial units, researchers can effectively explore urban environments across diverse spatial scales, thereby fostering a spatially coherent perception of these spaces in a more comprehensive environmental context. Moreover, the utilization of street-level imagery to study these units offers a unique opportunity to combine them with other data sources for multimodal learning. Multimodal learning allows for a more holistic comprehension of urban spaces, avoiding the isolation of street-level imagery from other important contextual data sources (Barbierato, Bernetti, Capecchi, & Saragosa, 2020; Ye, Zhang, Mu, Gao, & Liu, 2021).

In practice, the utilization of street-level imagery to represent urban spatial units presents a significant challenge due to the highly varying numbers of street-level images across different spatial units. Fig. 1 illustrates the distribution of street-level sample points in Shenzhen, revealing a significantly higher concentration of sample points in the downtown area relative to suburban regions. This is due to the nature of the street view sampling method, which follows the road network, which is typically denser and more complex in downtown regions than in suburban areas. In this case, there is a significant variation in the number of samples across spatial units, which would result in an imbalanced training set for deep learning models. To address this challenge, previous studies used pre-trained models to extract features from street-level images, and then employed numerical computing methods such as mean and maximum values to fuse these hidden features within a spatial unit (Verma, Jana, & Ramamritham, 2020). For example, Liu et al. (2021) used a pre-trained PSPNet to extract features from each street-level image, and subsequently averaged them as an environmental feature based on the unit. However, such approaches may fail to capture the semantic and spatial associations between street-level images within a spatial unit, which form local features that cannot be fully captured using numerical computing methods.

To address this challenge, this study proposes a novel vision long short-term memory (VisionLSTM) module to obtain vector representations from varying numbers of street-level images in spatial units. This module combines the advantages of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNN is utilized to extract the semantic features of each image as local features of spatial units, while RNN is applied to capture the associations among these local features. Inspired by natural language processing, this module considers the varying numbers of street-level images as an unordered sequence, allowing it to effectively extract overall features.

To validate the effectiveness of Vision-LSTM, this study conducts a series of experiments in Shenzhen to recognize urban villages, which are informal settlements resulting from rapid urbanization in China (Li & Wu, 2013). Due to their complex residential structures, urban villages exhibit intricate visual morphology and human activity patterns. Three types of multimodal information, including remote sensing imagery, street-level imagery, and social sensing data, are fused to accurately characterize urban villages and achieve reliable recognition results. The experiments demonstrate the efficacy of Vision-LSTM in capturing associations between street-level images and the benefits of multimodal learning at the spatial unit scale, which combines street-level imagery with other informative data. Such results enhance the research value of street-level imagery and facilitate multimodal learning-based urban research to deepen our understanding of urban spaces.

PROOF

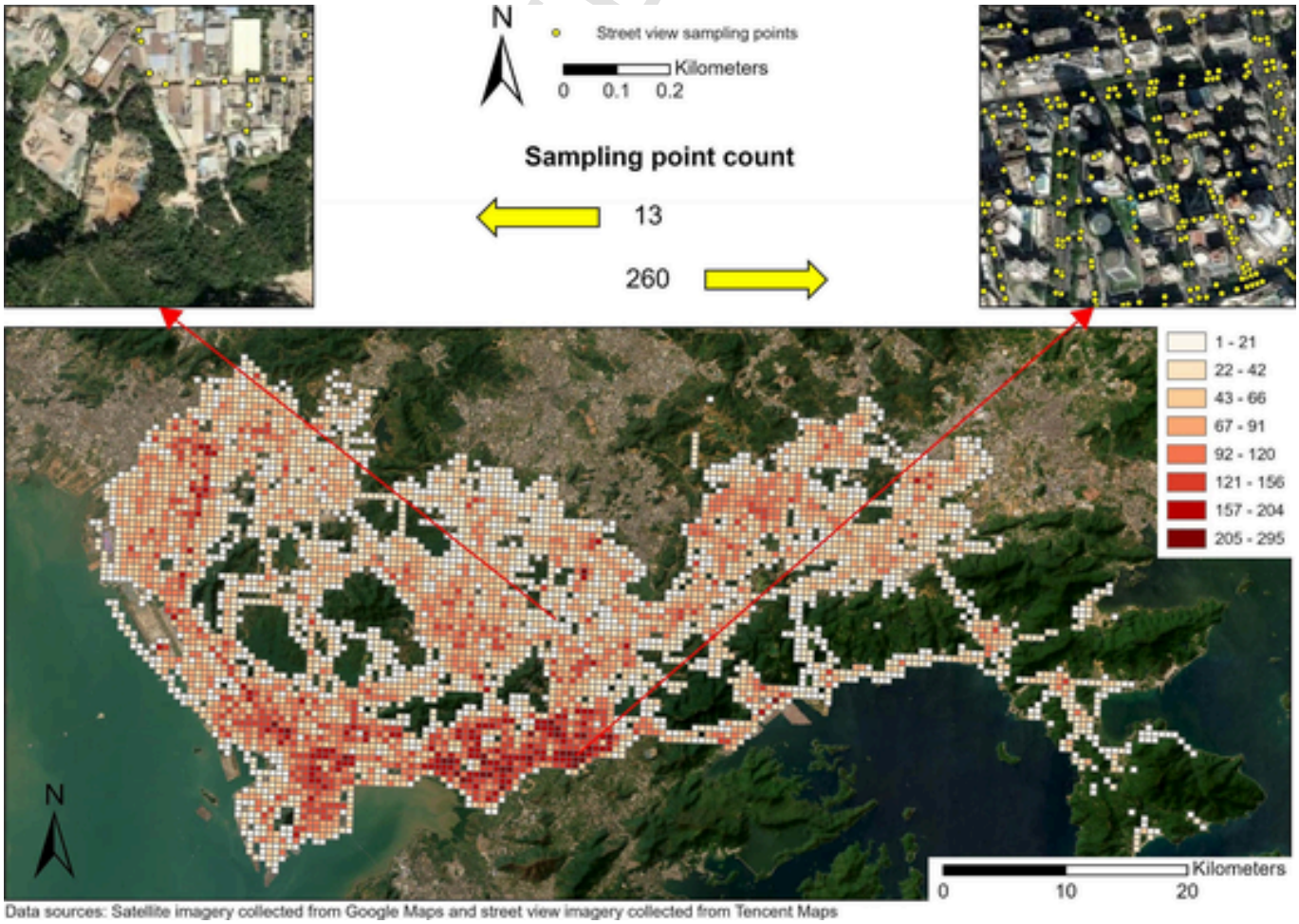


Fig. 1. Distribution of available street-level images in Shenzhen. There is a significant disparity in the number of samples across spatial units.

2. Related work

2.1. Using street-level imagery to understand urban spaces

Companies such as Google, Tencent, and Baidu now offer street view services that provide a detailed portrayal of urban spaces from a human perspective view. The street-level imagery collected by these companies comprehensively covers the road networks of most cities worldwide, offering an abundant data source for quantitative analysis of cities' physical visual environment (Biljecki & Ito, 2021). Street-level imagery has inherent advantages over traditional data sources, including easy access, high spatiotemporal coverage, and objective and standardized views of the built environment from embedded vantage points (Ibrahim et al., 2020; Kang, Zhang, Gao, Lin, & Liu, 2020; Rzotkiewicz, Pearson, Dougherty, Shortridge, & Wilson, 2018). Artificial intelligence techniques have facilitated the widespread use of street-level imagery in quantitative features of urban spaces. This enables researchers not only to represent and analyze physical space quantitaively but also to infer the socioeconomic status of urban areas (Helbich et al., 2019; Kang, Zhang, Gao, Peng, & Ratti, 2021; Liang, Zhao, & Biljecki, 2023; Sun, Zhang, Duarte, & Ratti, 2022).

Street-level imagery enables the spatial distribution of scene elements in physical urban spaces to be accessed. For example, an algorithm was developed by Doersch, Singh, Gupta, Sivic, and Efros (2012) that automatically distinguishes Paris from other cities using features like windows and balconies in street-level imagery. Using deep learning models, street-level imagery from numerous cities worldwide was analyzed to extract the percentage of vegetation, enabling a comparison of greenery across these urban areas (Li et al., 2015; Li & Ratti, 2018; Seiferling et al., 2017). Furthermore, by combining information such as shooting angles and geometric features of street-level imagery, it is possible to estimate sky openness under specific observation viewpoints (Gong et al., 2018; Ye, Zeng, Shen, Zhang, & Lu, 2019), the coverage area of solar radiation (Li & Ratti, 2018), and other factors. This can help to estimate the photovoltaic potential of the built-up area of the city (Li & Ratti, 2019; Liu et al., 2019) and areas where vehicle driving may cause solar glare (Li, Cai, Qiu, Zhao, & Ratti, 2019).

In addition to providing visible physical information about urban spaces, street-level imagery can also reveal hidden information related to residents' perceptions and socio-economic status. Zhang et al. (2018) utilized street-level imagery and deep learning to analyze the spatial distribution of human perceptions in Beijing and Shanghai, and further explored the "perception bias" between such perceptions and actual criminal data (Zhang, Fan, Kang, Hu, & Ratti, 2021). Furthermore, street-level imagery can also be used to predict house prices based on house photos and the surrounding environmental conditions (Kang et al., 2021). Street-level imagery has also been widely utilized in urban function recognition to uncover urban form (Cao et al., 2018; Gong, Ma, Kan, & Qi, 2019).

It is worth noting that most of these current studies focus only on the semantic representation of individual images, while ignoring the associations among multiple street-level images within a spatial unit. Linking the image semantics and spatial semantics of multiple street-level images in a spatial unit allows for an overall description of the urban spaces, thus allowing for a spatially coherent perception and supporting the analysis of urban space at different spatial scales (streets, blocks, etc.). Therefore, this study developed a deep learning module, Vision-LSTM, to address this issue.

2.2. Fusion approaches for representing regional features

In the context of fusing data from multiple locations to represent a regional feature, previous studies have predominantly relied on employing statistical methods. For instance, it's common practice to use points of interest (POIs) categories within a region as a means to represent its regional functionality (Yuan, Zheng, & Xie, 2012). Similarly, Kang et al. (2019) formulated emotional indices, such as the Joy Index and Average Happiness Index, to aggregate point-based emotion to place emotion. Regarding time-series data at the point level, Yao et al. (2022) employed a weighted version of the dynamic time warping barycentric averaging algorithm for the integration of time series of multiple buildings. This algorithm assigns different weights to different time series to preserve the time series' patterns.

Similar approaches are also common for the fusion of varying numbers of images. A prevalent strategy in this regard, particularly in deep learning neural networks, is to use CNNs to extract image features and then aggregate these features using modules such as average pooling, maximum pooling, or element-wise sum, as illustrated in Fig. 2 (Verma et al., 2020). For example, Liang et al. (2023) averaged the physical and perceptual features from the street-level images based on research units. However, these methods ignore the association between street-level images within the unit. Cao et al. (2018) extracted hidden features of street-level images with CNN. Subsequently, spatial interpolation was conducted within the unit for each hidden feature dimension, re-

Fig. 2. Average pooling, maximum pooling and element-wise sum for the fusion of street-level image features, in comparison of our proposed Vision-LSTM (Fig. 3).

CORRECTED PROOF

sulting in the utilization of a feature matrix to represent the unit's characteristics. Although this approach considers spatial correlation, the fact that spatial interpolation cannot serve as a module in a neural network makes it impossible to perform end-to-end learning. Moreover, for some popular feature extractors of street-view images, the hidden feature dimension ranges from 512 to 4196, leading to an exceedingly large feature matrix for each unit, thereby affecting computational efficiency (He, Zhang, Ren, & Sun, 2016).

While multi-view strategies in computer vision are valuebale for extracting features from multiple images, they are limited to a fixed number of images (Zhao, Xie, Xu, & Sun, 2017). In this study, the number of street-level images varies across spatial units ranging from several to hundreds due to the non-uniform distribution of sampling points. This variation in the number of images makes it challenging to transfer multi-view strategies to the fusion of varying numbers of street-level images. To address this issue, we propose the Vision-LSTM module, which draws on the experiences of applying natural language processing techniques to extract features from variable-length sequences. The Vision-LSTM module allows for the fusion of varying numbers of street-level images based on their semantic association, thus enabling the representation of regional features more accurately and effectively.

2.3. Using multimodal data to recognize urban villages

The growth of informal settlements (e.g., slums and shanty towns) is a global phenomenon accompanying increasingly urban sprawl. According to UN-Habitat (2013), an estimated 25% of the world's urban population lives in informal settlements. In China, informal settlements are commonly known as “urban villages” (Li & Wu, 2013). Urban villages are residential areas characterized by small, overcrowded, and self-built houses that lack security of tenure, basic services, and facilities (Huang, Liu, & Zhang, 2015). They emerge from the gradual disappearance of boundaries between urban and suburban rural areas (Brindley, 2003; Chung, 2010). The low living standards in urban villages pose a significant challenge to urban management, as they are associated with issues such as black-odorous water, spread of infectious diseases, and social segregation. Thus, many countries, including China, have initiated programs to upgrade or demolish urban villages in response to the UN-Habitat (2003) agenda for “cities without slums”. However, locating and mapping urban villages in practice is a challenging task due to the complex and dynamic nature of land use distribution (Guan, Wei, Lu, Dai, & Su, 2018).

Conventional approaches for recognizing urban villages rely on field surveys, which can be challenging to scale up due to their time-consuming and labor-intensive nature. Studies have shown the potential of satellite imagery for recognizing urban villages, as it can capture the physical environment from a nadir view (Hofmann & Bekkarnayeva, 2017; Huang et al., 2015; Mast, Wei, & Wurm, 2020). However, urban villages cannot be effectively recognized solely based on satellite imagery (Guan et al., 2018). Such information is insufficient for tracking urban villages as it is visually similar to other land use types, such as suburban villages and dense residential zones.

In recent years, multimodal learning has become an increasingly popular approach for urban function recognition, yielding more informative and precise results (Crooks et al., 2015; Feng et al., 2021; Gao, Janowicz, & Couclelis, 2017; Yao et al., 2022; Yuan et al., 2012). Compared with common urban functions, urban villages are more complex and have a higher semantic concept that is summarized by people's daily life experiences, thus requiring more diverse data sources for their recognition (Chen, Feng, et al., 2022,Chen et al., 2022). Existing research has demonstrated that street-level imagery can provide a complementary view of urban space with a human-like perspective, as opposed to a nadir view, and can capture specific visual characteristics of urban villages such as the multi-layered structure of building facades or the presence of compact and dirty streets (Chen, Feng, et al., 2022). Furthermore, human mobility data implicit in social sensing (Liu et al., 2015) can supplement the visual information for urban village recognition, as human travel behaviors differ across different land use types and urban villages (Pei et al., 2014). The complex residential structure of urban villages results in distinct travel patterns for their residents, providing a useful cue for recognition.

Existing literature primarily relies on remote sensing techniques for recognizing and mapping urban villages (Mast et al., 2020; Shi et al., 2020). Chen, Tu, et al. (2022) employed social sensing data, such as taxi trajectory data and POI data, along with remote sensing imagery, to achieve multimodal learning for urban village recognition. However, this approach neglects the substantial visual information that street-level imagery can offer. In another study by Chen, Feng, et al. (2022), street-level imagery and remote sensing imagery were integrated to recognize urban villages. Nevertheless, this study falls short in adequately addressing the challenge of utilizing street-level imagery to accurately represent regional features. It randomly selects a single image from the area, which is arbitrary and fails to provide a comprehensive visual representation of the urban environment. In this paper, a multimodal model, including remote sensing imagery, street-level imagery, and taxi trajectory data, is used to fuse these modalities and build a comprehensive understanding of recognizing urban villages.

3. Framework

3.1. Vision-LSTM

The proposed Vision-LSTM module, as depicted in Fig. 3, comprises a Convolutional Neural Network (CNN) with shared weights and a Recurrent Neural Network (RNN). Varying numbers of street-level images in spatial units can be treated as image sets of different lengths. To this end, we utilize the zero-padding method, commonly used in natural language processing, to process such variable-length street-level image sets, as shown in Fig. 4. Subsequently, each image within the image sets is fed individually to a CNN model with shared weights to extract the semantic features of individual images, and the input image sets are processed as sets of image features. Notably, any blank images added during zero padding are ignored and do not affect the model parameters during training.

The effectiveness of LSTM in processing variable-length time-series data has been well-established in previous studies (Wang, Du, & Wang, 2020). To leverage its capability, the Vision-LSTM module employs LSTM to capture the associations between image features extracted from long and variable-length sets of street-level images. However, since LSTM does not consider the 2dimensional spatial information, the image features in the sets are fed to the LSTM in a random order as time steps during training. The resulting deep features reflect the physical environment observed from the pedestrian's perspective and are indicative of overall features of spatial units, based on the varying lengths of street-level image sets.

3.2. Multimodal model

A multimodal deep neural network is proposed in this study, which utilizes a combination of satellite imagery, street-level imagery, and taxi trajectory data as inputs. The architecture of the proposed model, as depicted in Fig. 5, comprises of three branches, each responsible for extracting distinct features from the input modality. The ResNet18 backbone model is employed to extract the features of the nadir view in the satellite imagery branch. Meanwhile, the proposed Vision-LSTM module is used to extract visual features from varying numbers of street-level images in the street-level imagery branch. The LSTM fully convolutional networks (LSTM-FCN) backbone model is employed in the taxi trajectory branch to learn mobility features from travel volume time-series data based on taxi trajectories. Finally, the extracted fea-

CORRECTED PROOF

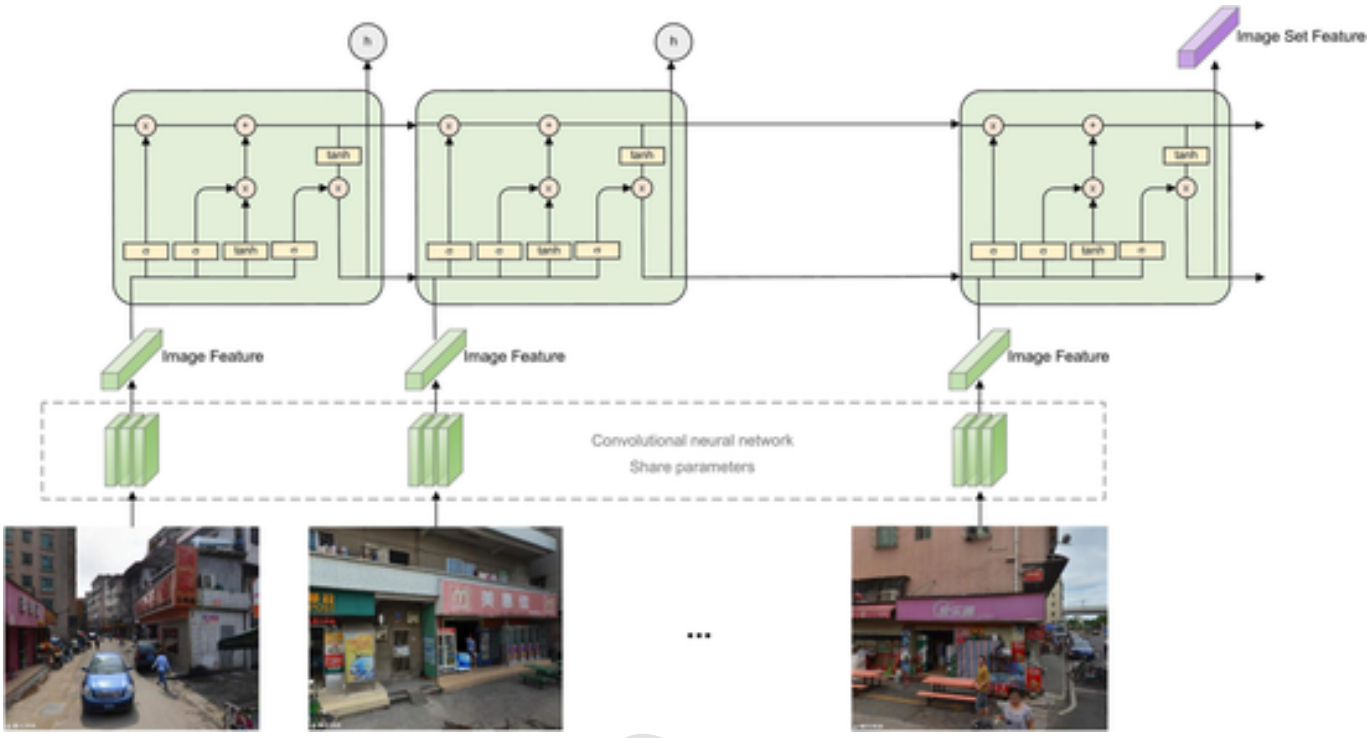


Fig. 3. Structure of the Vision-LSTM.

Fig. 4. Zero padding for street-level image sets.

tures from all the branches are concatenated and passed through a softmax layer to classify the urban and non-urban villages.

3.2.1. Satellite imagery branch

The satellite imagery branch serves to sense the physical environment of urban villages from a nadir view. To avoid overfitting issues arising from model complexity, ResNet18, a small and efficient CNN architecture, is adopted as a feature extractor for satellite imagery. ResNet, as proposed by He et al. (2016), is widely utilized to extract visual features from satellite images since it can learn rich feature representations of various scenes. The ResNet18 model parameters are pre-trained on over one million images from the ImageNet database (Krizhevsky, Sutskever, & Hinton, 2012) to classify images into 1000 object categories. However, since the texture of satellite imagery differs greatly from those in the ImageNet database, the model parameters are optimized and not frozen during training. To boost the model's generalization ability, the training set's satellite images are rotated and flipped for data augmentation.

Each satellite image is initially fed into a 7×7 convolutional layer and a 3×3 maximum pooling layer. Then, the resulting image feature tensors are passed through four residual blocks and consolidated into feature vectors by the average pooling layer. As a result, ResNet18 processes and summarizes each patch of the satellite image into a 512-dimensional numerical vector that captures semantic and contextual information about the physical environment from a nadir view.

3.2.2. Street-level imagery branch

The street-level imagery branch is designed to capture the physical environment from the perspective of pedestrians. To this end, the proposed Vision-LSTM module, discussed in Section 3.1, is utilized to extract the overall features of spatial units from varying numbers of street-level images. The image features are extracted by employing a ResNet18 model in the Vision-LSTM module, which is pre-trained on the Places 365 database (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017), a 10-million-image database designed for scene recognition, and better suited for street-level image classification. Given the complexity

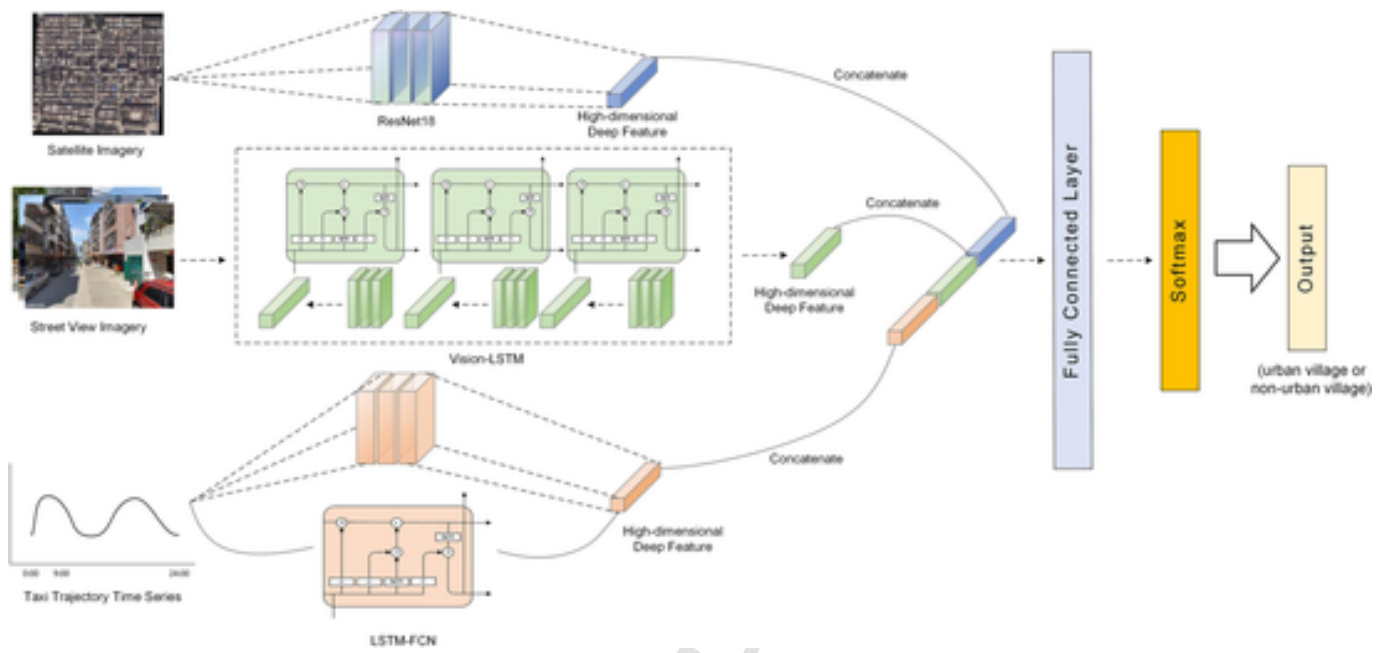


Fig. 5. Structure of multimodal model.

of the overall model, all parameters of the ResNet18 model in the street-level imagery branch are frozen during model training to prevent overfitting.

Once the ResNet18 model extracts the features of street-level images, each image is represented by a 512-dimensional feature. Next, these features are fused using LSTM to extract a overall 512-dimensional features of spatial units from varying numbers of street-level images. As with the satellite images, rotation and flip transformations are applied to street-level images to augment the training dataset.

3.2.3. Taxi trajectory branch

The aim of the taxi trajectory branch is to capture human mobility in spatial units using raw Global Positioning System (GPS) data from taxi trips. The GPS data is associated with unique car IDs and timestamps to recognize individual trips. The origin and destination (OD) of each trip are defined by the GPS coordinates of the point where the taxi status changes from vacant to occupied and from occupied to vacant, respectively. The time series of OD frequencies reflect taxi travel activity in spatial units, which is crucial for studying urban mobility patterns and functions (Mou, Cai, Zhang, Chen, & Zhu, 2019). The OD points are counted by spatial units and hourly timestamps to generate separate two time series for O and D, which are concatenated into a single long sequence. The sequence is then normalized to construct human mobility features based on the frequency of taxi trajectories.

This branch employs the LSTM-FCN model proposed by Karim, Majumdar, Darabi, and Chen (2018) to extract features from time series data on travel volume. The LSTM-FCN is well-suited for capturing temporal patterns of human activities and has been previously used for feature extraction of socioeconomic attributes (Yao et al., 2022). The LSTM-FCN model consists of two modules: the FCN module and LSTM module, which process the time-series data from two different perspectives. In the FCN module, the time-series data are treated as univariate time series with multiple time steps, and processed by three temporal convolution blocks for feature extraction. The feature vectors are then obtained by a global average pooling layer. The features extracted from the LSTM and FCN modules are concatenated and summarized to a 512-dimensional feature vector.

4. Experiments

4.1. Study area and data

Shenzhen, our study area, is located in the south of Guangdong Province, China, along the east coast of the Pearl River Delta (bounded by coordinates $113^{\circ}46' - 114^{\circ}37'E, 22^{\circ}27' - 22^{\circ}52'N$). As of the end of 2020, Shenzhen encompasses a total area of 1997.47 km^2 and consists of ten districts. With a permanent population of 17.63 million, it is the fourth most populous city in China. In the face of rapid urbanization, urban villages are increasingly prevalent in Shenzhen (Chen, Feng, et al., 2022; Lai, Jiang, & Xu, 2021; Wang, Wang, & Wu, 2009), as depicted in Fig. 6.

The multimodal data consist of high-resolution satellite imagery, street-level imagery, and taxi trajectory data.

- High-resolution satellite imagery for the year 2016 was collected from Google Earth, covering the entire Shenzhen city with a spatial resolution of 0.6 m per pixel and featuring three bands: red, green, and blue. To prepare the imagery for subsequent model input, it was uniformly segmented into subsets of 500 × 500 m.
- Street-level imagery was collected along the road network of Shenzhen in 2016 at intervals of 50 m using the Tencent street view application programming interface (API) (https://lbs.qq.com). A total of 292,037 sampling points were acquired, each containing a point ID, coordinates, and four street-level images captured from different angles, namely front, back, left, and right.
- Taxi trajectory data were collected from the smart GPS receiver installed inside taxis, which provide valuable insights into the mobility patterns of citizens (Wu et al., 2017). The data records information such as license plates, time, GPS coordinates, speed, and working status (occupied or vacant). Specifically, we collected a total of 860,436,489 records of taxi trajectories covering a 7-day period from October 23, 2017, to October 29, 2017, for our analysis.

We obtained the ground-truth data of urban villages from Chen, Tu, et al. (2022), which were generated through manual labeling based on satellite and street-level imagery, as well as official urban planning doc-

CORRECTED

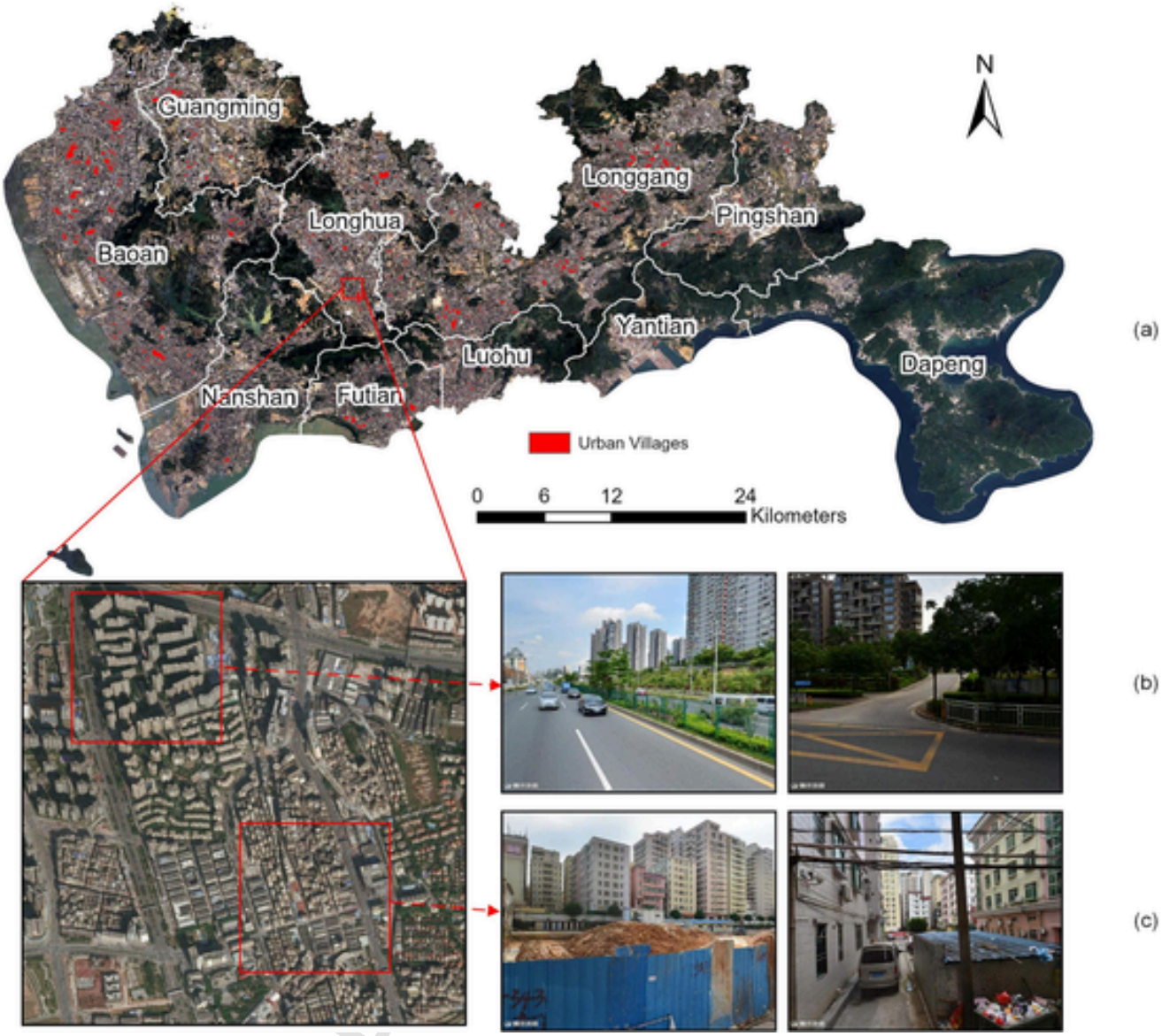


Fig. 6. Research area: Shenzhen, China. (a) Spatial distribution of urban villages and street-level images of (b) formal settlements and (c) urban villages are shown.

uments. The dataset provides the vector boundaries of each urban village within the city's expanse, and the spatial distribution of urban villages can be seen in Fig. 5. Given the stability of the built-up areas and planned demolitions of Shenzhen between 2016 and 2017, the temporal consistency of the multimodal data is not expected to significantly affect the recognition experiments. Nevertheless, it is worth noting that consistency over time would undoubtedly enhance the modeling and analysis of the data.

4.2. Experiment setup

This study adopts a grid-based approach with a spatial resolution of 500 m to construct the models. This selection ensures that each grid can be adequately represented with sufficient data. After filtering out the grids with missing data, the study obtained 4952 valid grids. Given that the excluded grids primarily include forest and lake parks, they do not significantly impact the recognition results. Each valid grid is characterized by three modalities, namely a satellite image, a varying number of street-level images, and taxi origin-destination (OD) time series (see Fig. 7). Refer to Table 1 for more details.

In this study, positive samples denote the grids that intersect with urban villages, while negative samples correspond to grids without any urban village areas. It is important to retain grids with a small proportion of urban village areas as positive samples because they are spatially proximate to urban villages, and are likely to provide valuable semantic information regarding urban villages in accordance with the first law of geography (Tobler, 1970). Determining the occupancy threshold for urban villages requires expert knowledge and experience, and is not transferable to other cities or spatial units of different resolutions. To ensure the model's validity and reliability, 20% of the samples were randomly selected as a validation dataset, while the remaining 80% of the samples were utilized for model training. The configuration details can be found in Table 2, and the spatial distribution of the samples is illustrated in Fig. 8.

Given the frozen parameters of the image feature extractor in the street-level imagery branch and the memory consumption of street-level images, the feature extraction process was conducted prior to model training. To optimize the model's performance, the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 was used, and the warm-up strategy (He et al., 2016) was adopted. Specifically, the learning rate was initially set to 0.0035 for 10 epochs to warm up

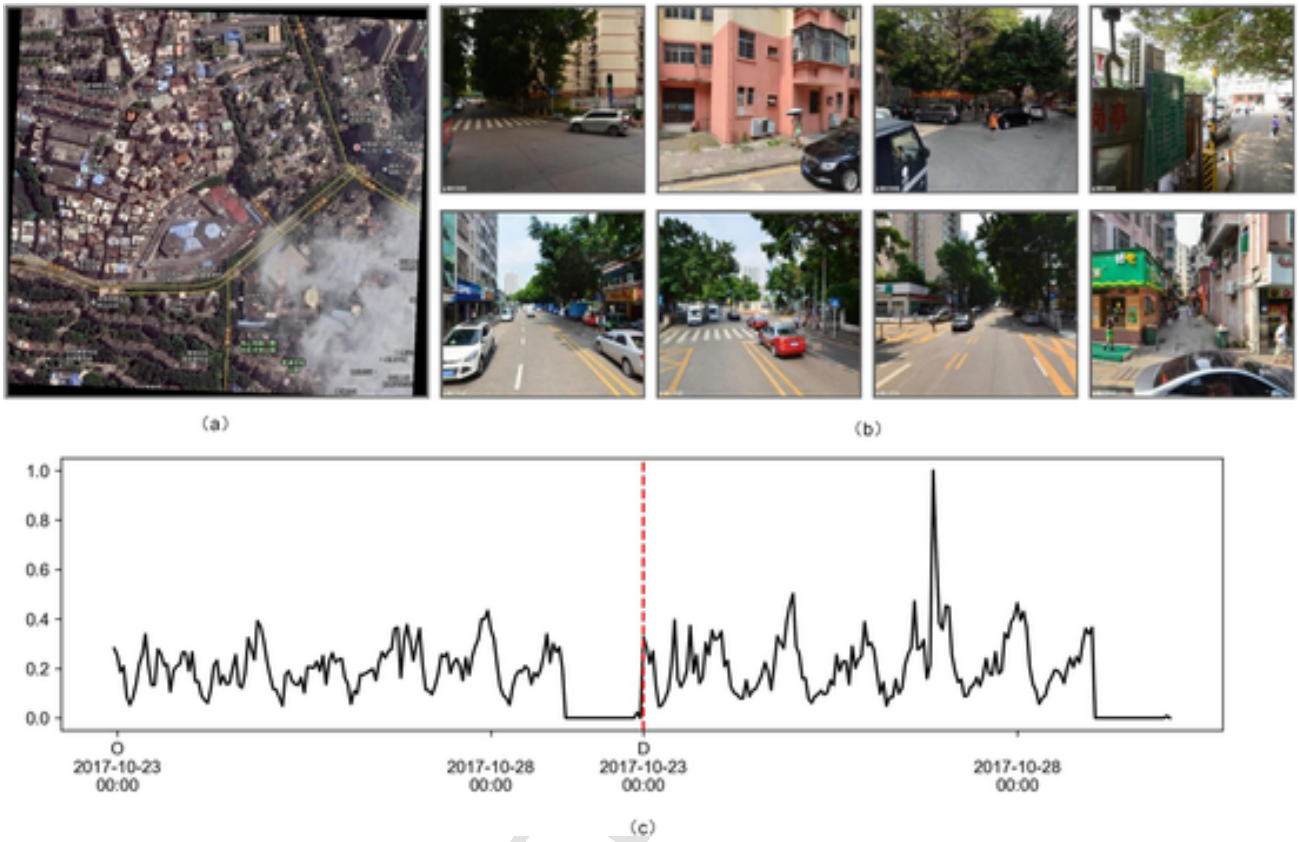


Fig. 7. Example from the dataset. This figure comprises three subfigures representing different data modalities: (a) satellite imagery, (b) street-level imagery, and (c) taxi trajectories. Due to the large number of street-level images, subfigure (b) displays a random selection of 8 images.

Table 1
Statistics of valid grids.

Statistical value	Number of street-level images	Travel volume
Mean	282	3633
Min	4	1
25%	108	39
Medium	232	232
75%	396	1845
Max	1692	160,803
Sum	1,396,791	17,992,053

Table 2
Configuration of sample numbers.

Data set	Positive sample	Negative sample
Training	3305	803
Validation	657	187

the training, after which it was reset to the initial value of 0.1 for the remainder of the training. The Cosine learning rate decay (Loshchilov & Hutter, 2017) was then used to determine the learning rate for each epoch, and the training process was continued for a total of 100 epochs with a batch size of 128. An early-stop method was employed to address the overfitting issue, whereby the training process would terminate if the validation loss failed to decrease after 10 epochs.

In view of the fact that the widely used cross-entropy loss may not be effective for unbalanced data sets, this study utilized the weighted cross-entropy (WCE) loss. The WCE loss can be calculated as:

$$WCEloss = -\sum_{i=1}^{n} w_i p_i \log\left(q_i\right) \quad (1)$$

where w_i represents the weight, and it is generally set to inversely proportional to their frequency in the training data set. p_i and q_i represent the true label and the predicted label, respectively.

To evaluate the model performance, the confusion matrix, overall accuracy (OA), Kappa index, and F1 score were adopted as evaluation metrics. All experiments were implemented on the PyTorch 1.10.0 framework with Python 3.7, and executed on a NVIDIA GeForce RTX 2080 Ti GPU with 11G memory and an Intel Xeon Gold 5118 @2.30GHz CPU. The operating system used was Ubuntu 18.04.

5. Results

5.1. Overall results

The present study yielded an OA of 92.0% on the validation dataset, accompanied by a Kappa index of 0.720 and an F1 score of 0.773. The distribution of correct confidence values for the samples, as illustrated in Fig. 9(a), indicates that approximately 20% of the samples exhibit a confidence level below 0.7. The median confidence level for all samples is found to be 0.897. The results suggest that the model performs well, as a majority of the samples are recognized with high confidence levels, which means that the features captured by the model make it easier to correctly identify most samples.

Fig. 9(b) displays the correlation between gross domestic product (GDP) and the mean confidence of administrative districts. The GDP data was collected from the Shenzhen statistical yearbook of 2017 (Bureau, 2017). The results indicate that highly developed or suburban districts tend to have higher confidence and are easier to recognize. For instance, Nanshan and Futian have high confidence (0.860 and 0.862, respectively) and high GDP values. In contrast, Dapeng and Guangming also exhibit high confidence (0.924 and 0.856, respectively), but their GDP values are low. Developing areas such as

CORRECTED

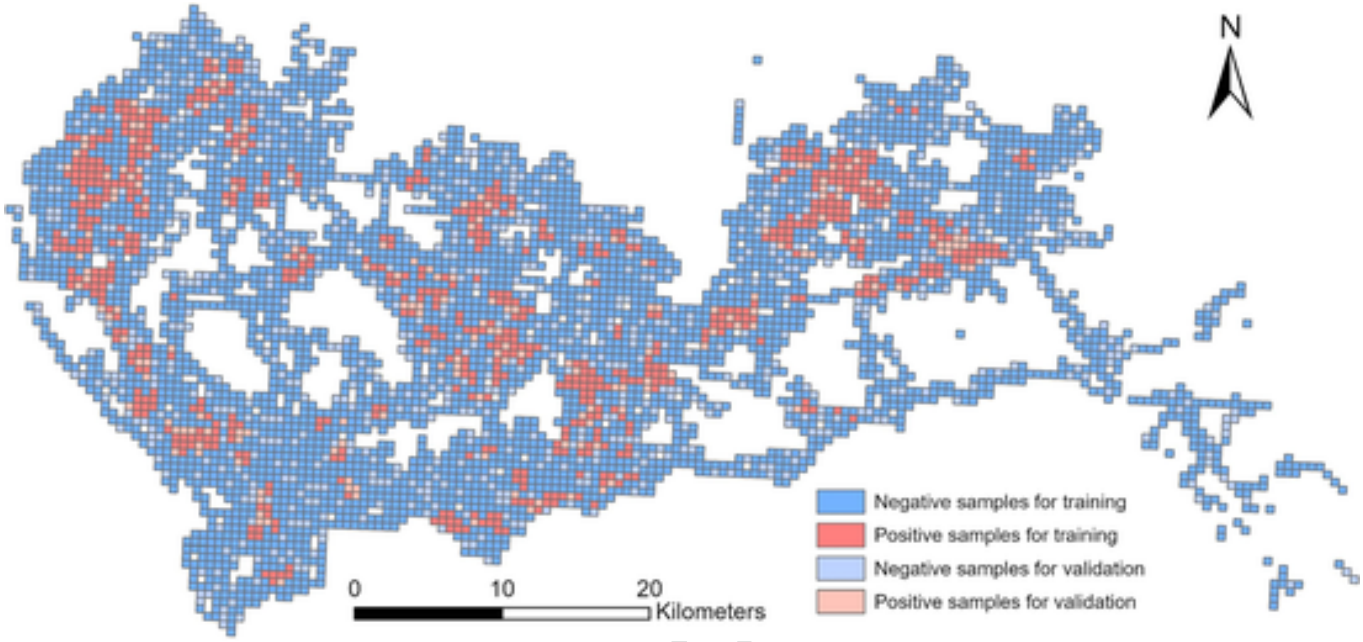


Fig. 8. Distribution of training and validation samples.

Luohu and Longhua have significantly lower confidence. Typically, factories and villages that have a high density of self-built houses are located in the rural-urban continuum and share similar characteristics with urban villages. Hence, in developing areas, the model may classify samples with low confidence that are easily confused with urban villages, which is reasonable.

Fig. 9(c) corroborates the findings presented in the previous paragraph. It is generally believed that the downtown of Shenzhen is the Nanshan-Futian area, which exhibits superior classification performance. As urbanization progresses outward, the confidence levels of adjacent districts begin to diminish. However, the remote areas show a resurgence in relatively high confidence levels.

Fig. 10 exhibits four different samples with varying degrees of confidence. Since the number of street-level images in samples is varying, the figure shows eight random street-level images of samples. The high-confidence urban village, as displayed in Fig. 10(a), is characterized by crowded streets with buildings situated in close proximity to each other. The mobility pattern of residents is not evident. Conversely, Fig. 10(b) portrays a typical formal settlement with clean streets and abundant vegetation. The travel activities of residents exhibit a clear daily pattern. Despite the overall good performance of the multimodal model, there are still certain challenging samples, such as cases where there is marked social segregation within grids, where it may fail to identify urban villages. Fig. 10(c) demonstrates this scenario, where although a significant part of the area is occupied by urban villages, the presence of non-urban village characteristics such as wide roads in most street-level images can misguide multimodal model. Similarly, formal villages with a dense distribution of self-built houses, as depicted in Fig. 10(d), may pose a difficulty in recognition.

5.2. Effectiveness of street-level imagery integration into multimodal learning

In order to analyze the effectiveness of street-level imagery integration into multimodal learning, we trained unimodal and bimodal models separately for comparison with the multimodal model in this study. To ensure fairness in the comparison, the same training and validation data sets as well as model hyperparameters were utilized for all models. The results of this analysis are presented in Table 3.

Results demonstrate that the model with Vision-LSTM fused with street-level images outperforms the other unimodal models, achieving an OA of 82.8%, Kappa of 0.540, and F1 of 0.647. These findings suggest that street-level images provide crucial information for recognizing urban villages. The unimodal model based on remote sensing imagery also performs well, with an OA of 81.8%, Kappa of 0.541, and F1 of 0.650. In contrast, the unimodal model based on taxi trajectories shows lower performance, with an OA of 71.9%. However, integrating taxi trajectory data with either satellite imagery or street-level imagery leads to significant improvement, with both models achieving an OA of approximately 87%. Notably, the bimodal model that combines satellite imagery and street-level imagery demonstrates the best performance, indicating that visual information plays a vital role in the recognition of urban villages, as these areas possess distinct visual and morphological characteristics.

Moreover, the results highlight the importance of multimodal learning in recognizing urban villages. When all three modalities are fused, the overall accuracy of the multimodal model significantly improves to 91.6%, which is a substantial improvement over the unimodal models. Additionally, the Kappa and F1 score gains are notable, with an increase of 0.180 and 0.126, respectively, compared to the unimodal model of street-level imagery. These results illustrate the significance of incorporating street-level imagery into multimodal learning for the recognition of urban villages.

6. Discussion

6.1. Fusing varying numbers of street-level images

Table 4 presents the results of four statistic fusion methods and our proposed Vision-LSTM module in two different resolution units. The stability of model significantly fluctuates when we do not use multiple image fusion and randomly select a street-level image to represent the spatial unit. As Table 4 demonstrates, for the 500-m grids, there is no significant improvement observed in no fusion and average pooling methods compared to the bimodal model that integrates remote sensing imagery and taxi trajectory data. The average pooling method improves the overall accuracy (OA) by 1.9% and Kappa and F1 by 0.046 and 0.040, respectively, compared to the bimodal model. However, the maximum pooling and element-wise sum methods per-

CORRECTED

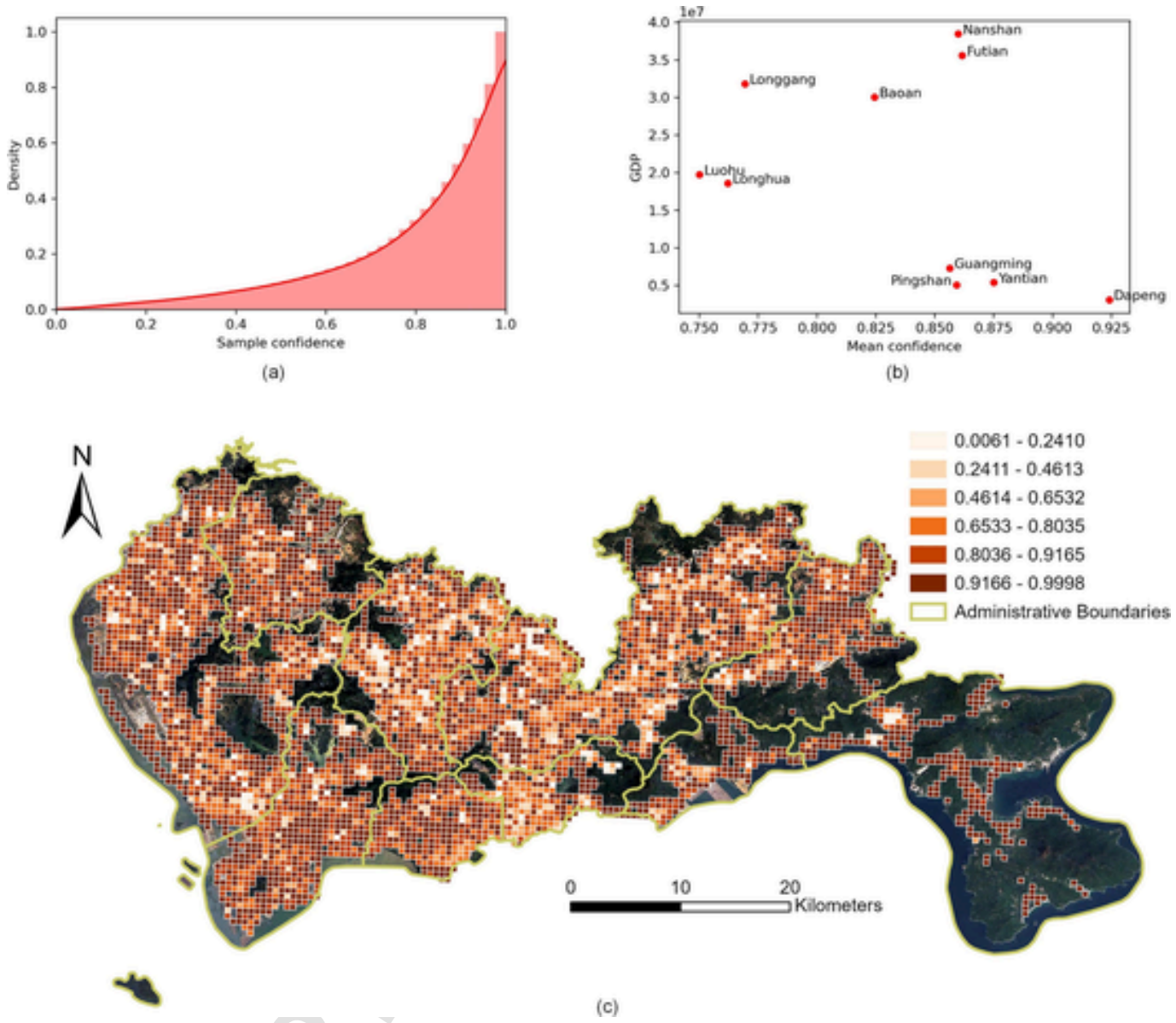


Fig. 9. Distribution of the confidence of correct label. (a) Cumulative distribution. (b) Scatter between gross domestic product (GDP) and mean confidence of each administrative district. (c) Spatial distribution. The GDP data was collected from Shenzhen statistical yearbook of 2017 (Bureau, 2017).

form worse than the bimodal model, possibly due to their inability to extract critical information, resulting in a reduction of differences between images. Notably, the proposed Vision-LSTM method shows a marked improvement with an OA and Kappa improved by 2.5% and 0.064, respectively, compared to the average pooling method.

Similar patterns exist in the models trained on the 250-m grids, whereas all the models trained on 500-m grids exhibit significantly superior performance compared to the models trained on 250-m grids. This disparity may arise due to the limited ability of the model to capture only local features of urban villages when the grid size is small. Consequently, these local features can be easily confused with other types of urban functions. Conversely, employing a 500-m grid enables models to capture a more comprehensive view of urban villages since the size of a 500-m grid is almost the size of a small urban village.

These findings suggest that LSTM can extract effective information from both image feature sets and long-time series. The feature spaces generated by statistic fusion methods are less efficient in distinguishing between urban villages and non-urban villages, while the Vision-LSTM module utilizes a parameter learning approach to extract dense features containing essential information about urban villages. Additionally, the number of parameters required for the Vision-LSTM module is small, and most of them can be fine-tuned by other classical models, which minimizes the computational resources and complexity of the multimodal model.

It is important to acknowledge that while Vision-LSTM is implemented within grid systems for the urban village case, its applicability is not confined to specific spatial units. VisionLSTM offers a versatile approach for effectively fusing multiple street-level images. In various contexts, it can be appropriately adapted to accommodate diverse unit systems, such as street blocks or H3 hexagons.

6.2. *Using multimodal data to understand urban functions*

With the increasing availability of sensing technology, researchers now have a greater range of data options that can reveal hidden information about cities. Different modalities of data provide distinct perspectives on urban features and functions and enable both humans and machines to gain a better understanding of urban spaces. For example,

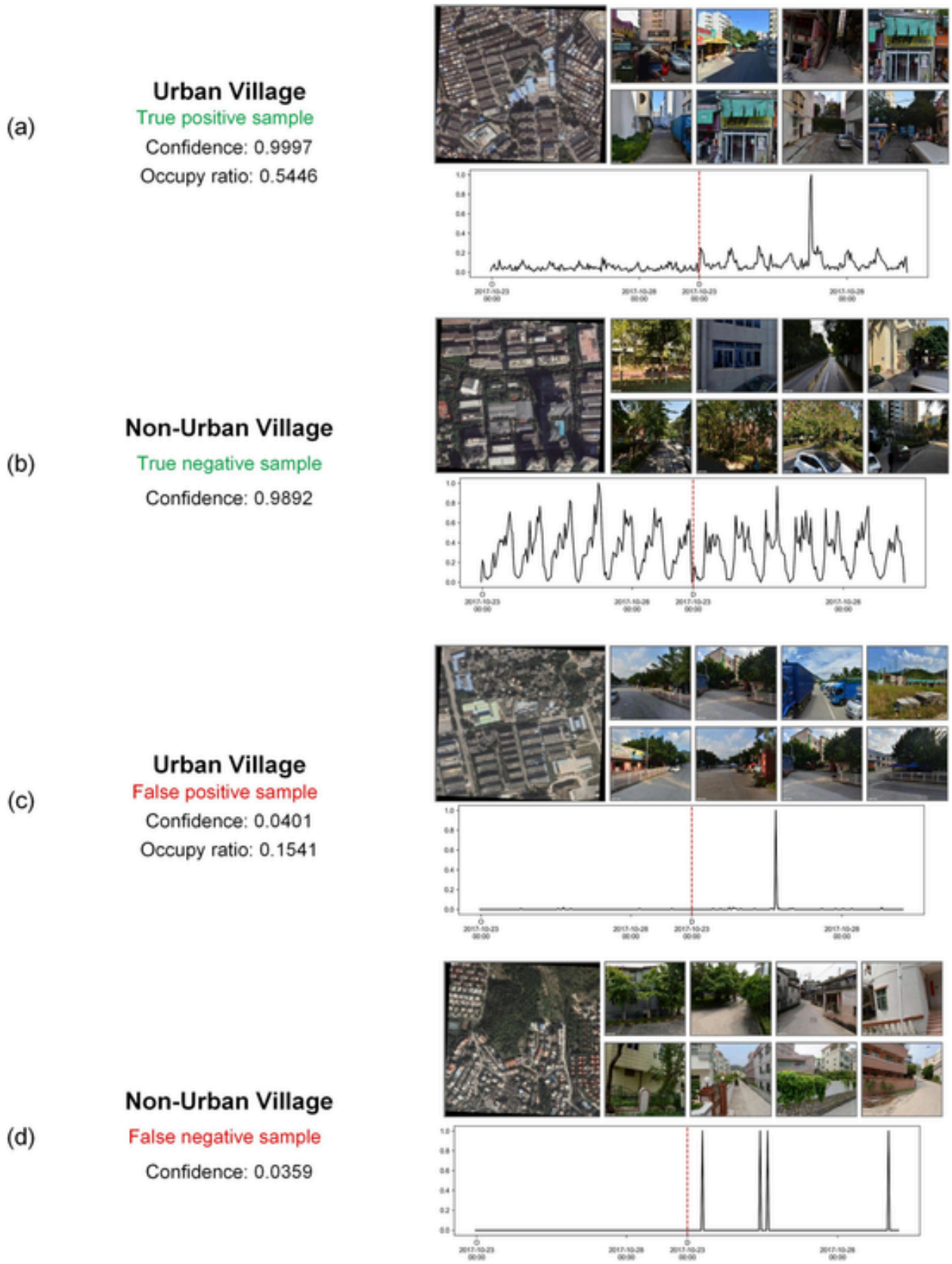


Fig. 10. Several samples in different confidence. (a) True positive urban village sample in high confidence. (b) True negative non-urban village sample in high confidence. (c) False negative urban village sample in low confidence. (d) False positive non-urban village sample in low confidence.

Table 3
Comparison of model assessment results trained by unimodal, bimodal and multimodal models. (OA: overall accuracy, SI: satellite imagery, SLI: street-level imagery).

Data	Model	OA (%)	Kappa	F1
SI	ResNet18	81.8	0.541	0.650
SLI	Vision-LSTM (proposed in this study)	82.8	0.540	0.647
Taxi trajectory	LSTM-FCN	71.9	0.311	0.477
SI + SLI	ResNet18 + Vision-LSTM	89.0	0.647	0.715
SI + Taxi trajectory	ResNet18 + LSTM-FCN	87.8	0.645	0.721
SLI + Taxi trajectory	Vision-LSTM + LSTM-FCN	87.2	0.588	0.668
SI + SLI + Taxi trajectory	ResNet18 + Vision-LSTM + LSTM-FCN	91.6	0.720	0.773

Table 4
Comparison of model assessment results trained by different fusion methods for varying numbers of street-level images. (OA: overall accuracy).

Resolution	Method	OA(%)	Kappa	F1
250 m	No fusion (random image)	79.7	0.382	0.484
	Average Pooling	77.0	0.359	0.471
	Maximum Pooling	72.5	0.311	0.436
	Element-wise Sum	69.1	0.273	0.411
	Vision-LSTM (proposed in this study)	80.5	0.407	0.491
500 m	No fusion (random image)	88.1	0.634	0.708
	Average Pooling	89.1	0.656	0.727
	Maximum Pooling	79.3	0.461	0.588
	Element-wise Sum	77.4	0.432	0.566
	Vision-LSTM (proposed in this study)	91.6	0.720	0.773

satellite imagery can provide information on surface entities and textures, while social sensing data like taxi trajectories can shed light on resident mobility patterns and social functions. Special urban functions, such as informal settlements and urban villages, are typically difficult to recognize using unimodal data, and may require expert knowledge to identify even through manual recognition (Mast et al., 2020). Consequently, intelligent processing and fusion of multimodal data are essential for the accurate recognition of urban functions. Street-level imagery used to difficult to combine with other data for multimodal learning although it has proven to be a powerful kind of data to understand urban spaces. With Vision-LSTM, researchers can use street-level imagery to depict the urban environment at various spatial scales, which sheds light on future urban studies.

Although the proposed multimodal model is used to recognize urban villages in this study, it can be applied to the recognition of other urban functions as well. The model focuses on sensing special information related to particular urban functions from multimodal data, rather than being restricted to the application of a specific type of urban function. Moreover, using a wider range of data from different modalities will likely improve the model's performance. The data sources used in this study, such as satellite imagery and street-level imagery, are relatively easy to obtain, while data such as taxi trajectory data may be more challenging to acquire. However, since taxi trajectory data serves as a proxy for human dynamics, other human activity data, such as check-in data (Liu, Sui, Kang, & Gao, 2014) and SafeGraph data (Chen, Bowers, Zhu, Gao, & Cheng, 2022), can be used as substitutes.

7. Conclusion

This study presents a novel deep learning module that integrates varying number of street-level images to represent the characteristics of an urban spatial unit. The proposed module is utilized in the context of recognizing urban villages in Shenzhen, and its integration into a multimodal model leads to a 91.6% accuracy. The results demonstrate that the module outperforms common statistical methods in capturing the complex connections between street-level images. By using this module, street-level imagery can be used more widely, such as urban studies at different spatial scales and multimodal learning by fusing other data. This approach has the potential to enhance the understanding of urban spaces, and can be applied to other urban contexts beyond the identification of urban villages.

Declaration of Competing Interest

To whom it may concern: We declare that none of the authors have competing financial or non-financial interests as defined by Computers, Environment and Urban Systems, Elsevier.

Acknowledgements

This work was sponsored by the National Natural Science Foundation of China under Grant 42371468 and CCF-Tencent Open Fund (RAGR20210101).

References

Barbierato, E., Bernetti, I., Capecchi, I., & Saragosa, C. (2020). *Integrating remote sensing and street view images to quantify urban forest ecosystem services*. *Remote Sensing*, *12*(2), 329.
Biljecki, F., & Ito, K. (2021). *Street view imagery in urban analytics and GIS: A review*. *Landscape and Urban Planning*, *215*, 104217.
Brindley, T. (2003). *The social dimension of the urban village: A comparison of models for sustainable urban development*. *Urban Design International*, *8*(1–2), 53–65.
Bureau, S.S. (2017). *Shenzhen statistic yearbook of 2017*. Shenzhen: China Statistics Press.
Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., … Qiu, G. (2018). *Integrating aerial and street view images for urban land use classification*. *Remote Sensing*, *10*(10), 1553.
Chen, B., Feng, Q., Niu, B., Yan, F., Gao, B., Yang, J., … Liu, J. (2022). *Multimodal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network*. *International Journal of Applied Earth Observation and Geoinformation*, *109*, 102794.
Chen, D., Tu, W., Cao, R., Zhang, Y., He, B., Wang, C., … Li, Q. (2022). *A hierarchical approach for fine-grained urban villages recognition fusing remote and social sensing data*. *International Journal of Applied Earth Observation and Geoinformation*, *106*, 102661.
Chen, T., Bowers, K., Zhu, D., Gao, X., & Cheng, T. (2022). *Spatio-temporal stratified associations between urban human activities and crime patterns: A case study in San Francisco around the COVID-19 stay-at-home mandate*. *Computational Urban Science*, *2*(1), 13.
Chung, H. (2010). *Building an image of villages-in-the-city: A clarification of China's distinct urban spaces: Debates and developments*. *International Journal of Urban and Regional Research*, *34*(2), 421–437.
Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., … Lamprianidis, G. (2015). *Crowdsourcing urban form and function*. *International Journal of Geographical Information Science*, *29*(5), 720–741.
Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A.A. (2012). *What makes Paris look like Paris? ACM Transactions on Graphics*, *31*(4), 1–9.
Duarte, F., & Ratti, C. (2021). What urban cameras reveal about the City: The work of the Senseable City lab. In W., Shi, M.F., Goodchild, M., Batty, M.-P., Kwan, & A., Zhang (Eds.), *Urban informatics* (pp. 491–502). Singapore: Springer Singapore.
Fan, Z., Zhang, F., Loo, B.P.Y., & Ratti, C. (2023). *Urban visual intelligence: Uncovering hidden city profiles with street view images*. *Proceedings of the National Academy of Sciences*, *120*(27), e2220417120.
Feng, Y., Huang, Z., Wang, Y., Wan, L., Liu, Y., Zhang, Y., & Shan, X. (2021). *An SOEBased learning framework using multisource big data for identifying urban functional zones*. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 7336–7348.
Gao, S., Janowicz, K., & Couclelis, H. (2017). *Extracting urban functional regions from points of interest and human activities on location-based social networks*. *Transactions in GIS*, *21*(3), 446–467.
Gong, F.-Y., Zeng, Z.-C., Zhang, F., Li, X., Ng, E., & Norford, L.K. (2018). *Mapping sky, tree, and building view factors of street canyons in a high-density urban environment*. *Building and Environment*, *134*, 155–167.
Gong, Z., Ma, Q., Kan, C., & Qi, Q. (2019). *Classifying street spaces with street view images for a spatial Indicator of urban functions*. *Sustainability*, *11*(22), 6424.
Guan, X., Wei, H., Lu, S., Dai, Q., & Su, H. (2018). *Assessment on the urbanization strategy in China: Achievements, challenges and reflections*. *Habitat International*, *71*, 97–109.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).
Helbich, M., Yao, Y., Liu, Y., Zhang, J., Liu, P., & Wang, R. (2019). *Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China*. *Environment International*, *126*, 107–117.
Hofmann, P., & Bekkarnayeva, G. (2017). Object-based change detection of informal settlements. *In 2017 Joint urban remote sensing event (JURSE)*. New York: IEEE.
Huang, X., Liu, H., & Zhang, L. (2015). *Spatiotemporal detection and analysis of urban*

CORRECTED PROOF

villages in mega city regions of China using high-resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(7), 3639–3657.
Huang, X., Yang, J., Li, J., & Wen, D. (2021). *Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery*. *ISPRS Journal of Photogrammetry and Remote Sensing*, *175*, 403–415.
Ibrahim, M.R., Haworth, J., & Cheng, T. (2020). *Understanding cities with machine eyes: A review of deep computer vision in urban analytics*. *Cities*, *96*, 102481.
Kang, Y., Jia, Q., Gao, S., Zeng, X., Wang, Y., Angsuesser, S., … Fei, T. (2019). *Extracting human emotions at different places based on facial expressions and spatial clustering analysis*. *Transactions in GIS*, *23*(3), 450–480.
Kang, Y., Zhang, F., Gao, S., Lin, H., & Liu, Y. (2020). *A review of urban physical environment sensing using street view imagery in public health studies*. *Annals of GIS*, *26*(3), 261–275.
Kang, Y., Zhang, F., Gao, S., Peng, W., & Ratti, C. (2021). *Human settlement value assessment from a place perspective: Considering human dynamics and perceptions in house price modeling*. *Cities*, *118*, 103333.
Karim, F., Majumdar, S., Darabi, H., & Chen, S. (2018). *LSTM fully convolutional networks for time series classification*. *IEEE Access*, *6*, 1662–1669.
Khosla, A., An An, B., Lim, J.J., & Torralba, A. (2014). Looking beyond the visible scene. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3710–3717).
Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *In Proceedings of the 25th international conference on neural information processing systems - Volume 1, NIPS'12, pages 1097–1105*. Red Hook, NY: USA. Curran Associates Inc.
Lai, Y., Jiang, L., & Xu, X. (2021). *Exploring Spatio-temporal patterns of Urban Village redevelopment: The case of Shenzhen, China*. *Land*, *10*(9), 976.
Li, X., Cai, B.Y., Qiu, W., Zhao, J., & Ratti, C. (2019). *A novel method for predicting and mapping the occurrence of sun glare using Google street view*. *Transportation Research Part C: Emerging Technologies*, *106*, 132–144.
Li, X., & Ratti, C. (2018). *Mapping the spatial distribution of shade provision of street trees in Boston using Google street view panoramas*. *Urban Forestry & Urban Greening*, *31*, 109–119.
Li, X., & Ratti, C. (2019). *Mapping the spatio-temporal distribution of solar radiation within street canyons of Boston using Google street view panoramas and building height model*. *Landscape and Urban Planning*, *191*, 103387.
Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). *Assessing street-level urban greenery using Google street view and a modified green view index*. *Urban Forestry & Urban Greening*, *14*(3), 675–685.
Li, Z., & Wu, F. (2013). *Residential satisfaction in China's informal settlements: A case study of Beijing, Shanghai, and Guangzhou*. *Urban Geography*, *34*(7), 923–949.
Liang, X., Zhao, T., & Biljecki, F. (2023). *Revealing spatio-temporal evolution of urban visual environments with street view imagery*. *Landscape and Urban Planning*, *237*, 104802.
Liu, K., Yin, L., Zhang, M., Kang, M., Deng, A.-P., Li, Q.-L., & Song, T. (2021). *Facilitating fine-grained intra-urban dengue forecasting by integrating urban environments measured from street-view images*. *Infectious Diseases of Poverty*, *10*(1), 40.
Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). *Social sensing: A new approach to understanding our socioeconomic environments*. *Annals of the Association of American Geographers*, *105*(3), 512–530.
Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). *Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data*. *PLoS One*, *9*(1), e86026.
Liu, Z., Yang, A., Gao, M., Jiang, H., Kang, Y., Zhang, F., & Fei, T. (2019). *Towards feasibility of photovoltaic road for urban traffic-solar energy estimation using street view image*. *Journal of Cleaner Production*, *228*, 303–318.
Loshchilov, I., & Hutter, F. (2017). *SGDR: Stochastic gradient descent with warm restarts*.
Mast, J., Wei, C., & Wurm, M. (2020). *Mapping urban villages using fully convolutional neural networks*. *Remote Sensing Letters*, *11*(7), 630–639.
Mou, X., Cai, F., Zhang, X., Chen, J., & Zhu, R. (2019). Urban function identification based on POI and taxi trajectory data. *Proceedings of the 2019 3rd international conference on big data research, pages 152–156, Cergy-Pontoise France*. ACM.
Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). *A new insight into land use classification based on aggregated mobile phone data*. *International Journal of Geographical Information Science*, *28*(9), 1988–2007.
Rzotkiewicz, A., Pearson, A.L., Dougherty, B.V., Shortridge, A., & Wilson, N. (2018). *Systematic review of the use of Google street view in health research: Major themes, strengths, weaknesses and possibilities for future research*. *Health & Place*, *52*, 240–246.
Seiferling, I., Naik, N., Ratti, C., & Proulx, R. (2017). *Green streets - quantifying and mapping urban trees with street-level imagery and computer vision*. *Landscape and Urban Planning*, *165*, 93–101.
Shi, Q., Liu, M., Liu, X., Liu, P., Zhang, P., Yang, J., & Li, X. (2020). *Domain adaption for fine-grained Urban Village extraction from satellite images*. *IEEE Geoscience and Remote Sensing Letters*, *17*(8), 1430–1434.
Sun, M., Zhang, F., Duarte, F., & Ratti, C. (2022). *Understanding architecture age and style through deep learning*. *Cities*, *103787*.
Tobler, W.R. (1970). *A computer movie simulating urban growth in the Detroit Region*. *Economic Geography*, *46*(sup1), 234–240.
UN-Habitat (Ed.). (2003). The challenge of slums: Global report on human settlements. *Earthscan publications, London*. Sterling: VA.
UN-Habitat (2013). *Streets as public spaces and drivers of urban prosperity*. UN-HABITAT.
Verma, D., Jana, A., & Ramamritham, K. (2020). *Predicting human perception of the urban environment in a spatiotemporal urban setting using locally acquired street view images and audio clips*. *Building and Environment*, *186*, 107340.
Wang, J.Q., Du, Y., & Wang, J. (2020). *LSTM based long-term energy consumption prediction with periodicity*. *Energy*, *197*, 117197.
Wang, Y.P., Wang, Y., & Wu, J. (2009). *Urbanization and informal development in China: Urban villages in Shenzhen*. *International Journal of Urban and Regional Research*, *33*(4), 957–973.
Wu, L., Hu, S., Yin, L., Wang, Y., Chen, Z., Guo, M., … Xie, Z. (2017). *Optimizing cruising routes for taxi drivers using a spatio-temporal trajectory model*. *ISPRS International Journal of Geo-Information*, *6*(11), 373.
Yao, Y., Yan, X., Luo, P., Liang, Y., Ren, S., Hu, Y., … Guan, Q. (2022). *Classifying land-use patterns by integrating time-series electricity data and high-spatial resolution remote sensing imagery*. *International Journal of Applied Earth Observation and Geoinformation*, *106*, 102664.
Ye, C., Zhang, F., Mu, L., Gao, Y., & Liu, Y. (2021). *Urban function recognition by integrating social media and street-level imagery*. *Environment and Planning B: Urban Analytics and City Science*, *48*(6), 1430–1444.
Ye, Y., Zeng, W., Shen, Q., Zhang, X., & Lu, Y. (2019). *The visual quality of streets: A human-centred continuous measurement based on machine learning algorithms and street view images*. *Environment and Planning B: Urban Analytics and City Science*, *46*(8), 1439–1457.
Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. *In proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '12, page 186*. Beijing: China. ACM Press.
Zhang, F., Fan, Z., Kang, Y., Hu, Y., & Ratti, C. (2021). *"Perception bias": Deciphering a mismatch between urban crime and perception of safety*. *Landscape and Urban Planning*, *207*, 104003.
Zhang, F., Wu, L., Zhu, D., & Liu, Y. (2019). *Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns*. *ISPRS Journal of Photogrammetry and Remote Sensing*, *153*, 48–58.
Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H., & Ratti, C. (2018). *Measuring human perceptions of a large-scale urban region using machine learning*. *Landscape and Urban Planning*, *180*, 148–160.
Zhang, F., Zhou, B., Ratti, C., & Liu, Y. (2019). *Discovering place-informative scenes and objects using social media photos*. *Royal Society Open Science*, *6*(3), 181375.
Zhao, J., Xie, X., Xu, X., & Sun, S. (2017). *Multi-view learning overview: Recent progress and new challenges*. *Information Fusion*, *38*, 43–54.
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). *Places: A 10 million image database for scene recognition*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.

CORRECTED PROOF