This is the Accepted Manuscript version of an article published by Elsevier in the journal *Computers, Environment and Urban Systems* in 2023, which is available at:
https://doi.org/10.1016/j.compenvurbsys.2022.101915
Cite as: Zhao T, Liang X, Tu W, Huang Z, Biljecki F (2023): Sensing urban soundscapes from street view imagery. *Computers, Environment and Urban Systems*, 99: 101915.

Sensing urban soundscapes from street view imagery

Tianhong Zhao[a,b], Xiucheng Liang[b], Wei Tu[a], Zhengdong Huang[a], Filip Biljecki[b,c,*]

[a] *School of Architecture and Urban Planning, Shenzhen University, China*
[b] *Department of Architecture, National University of Singapore, Singapore*
[c] *Department of Real Estate, National University of Singapore, Singapore*

Abstract

A healthy acoustic environment is an essential component of sustainable cities. Various noise monitoring and simulation techniques have been developed to measure and evaluate urban sounds. However, sensing large areas at a fine resolution remains a great challenge. Based on machine learning, we introduce a new application of street view imagery — estimating large-area high-resolution urban soundscapes, investigating the premise that we can predict and characterize soundscapes without laborious and expensive noise measurements. First, visual features are extracted from street-level imagery using computer vision. Second, fifteen soundscape indicators are identified and a survey was conducted to gauge them solely from images. Finally, a prediction model is constructed to infer the urban soundscape by modeling the non-linear relationship between them. The results are verified with extensive field surveys. Experiments conducted in Singapore and Shenzhen using half a million images affirm that street view imagery enables us to sense large-scale urban soundscapes with low cost but high accuracy and detail, and provides an alternative means to generate soundscape maps. R^2 reaches 0.48 by evaluating the predicted results with field data collection. Further novelties in this domain are revealing the contributing visual elements and spatial laws of soundscapes, underscoring the usability of crowdsourced data, and exposing international patterns in perception.

Keywords: urban planning, GeoAI, perception, spatial analysis, deep learning, built environment

1. Introduction

Perception of urban environment is an essential task in urban informatics (Zhang et al., 2018, 2021a; Kruse et al., 2021), as it relates to urban design and planning (De Silva et al., 2017), public health (Harvey et al., 2015), and living quality (Van Renterghem et al., 2020; Kang et al., 2021). The acoustic environment is a critical component of the urban environment due to the direct impact on physical and mental health, e.g. a bad acoustic environment increases the risk of hypertension and heart attack (Stansfeld et al., 2005; Hoffmann et al., 2006) while pleasant sounds promote public health (Andringa and Lanser, 2013). Traditional assessments of the acoustic environment rely on the use of sound level meters, which describe sound in decibels (dB). Such an assessment focuses only on the negative aspects of sound but ignores the fact that some sounds (e.g., nature and bird sounds, etc.) have a positive impact on people's health (Nilsson and Berglund, 2006). The soundscape is proposed to involve how the acoustic environment affects the human perception of cities. According to the widely accepted definition given in the International Standard ISO, the soundscape is 'acoustic environment as perceived or experienced and/or understood by a person or people, in context' (ISO/DIS 12913-1, 2014). This concept represents a paradigm shift in the field of acoustic environment evaluation, as it focuses on human perception rather than physical measurements. (Brooks et al., 2014; Hasegawa and Lau, 2022). Sensing soundscape helps to improve the perceived quality of the acoustic environment and — as a result — plays an essential role in health betterment.

A variety of research for sensing and evaluating soundscapes has been proposed, thereby improving the quality of soundscapes, e.g. placing noise sensors in locations such as airports and construction sites. However, these solutions are costly and cover a rather small area, inhibiting such implementation at the city-scale. Recently, researchers have begun to develop methods and multi-source data for assessing soundscapes that are both cheap and large-scale (Hsieh et al., 2015; Verma et al., 2019; Gasco et al., 2020). For example, Becker et al. (2013) proposed that patients participate in crowdsourcing via a smartphone app that collects and measures the noise samples and subjective (opinions, feelings) data. Aiello et al. (2016) combined social media data with geo-referenced images and text, analyzed sound-related words using text mining techniques, and evaluated the soundscape distribution of London and Barcelona. While these methods address the issues of

*Corresponding author

Email addresses: zhaotianhong2016@email.szu.edu.cn (Tianhong Zhao), xiucheng@u.nus.edu (Xiucheng Liang), tuwei@szu.edu.cn (Wei Tu), zdhuang@szu.edu.cn (Zhengdong Huang), filip@nus.edu.sg (Filip Biljecki)

costly and small-scale to some extent, the difficulties in controlling the quality of these data suggest that new data sources for soundscape assessment be investigated.

Our hypothesis, investigated in this paper, is that Street View Imagery (SVI) is a valuable data source to evaluate soundscapes, as the human visual and auditory perception is inextricably linked, which was proven by psychological experiments (Salem et al., 2018; Verma et al., 2020; Einhäuser et al., 2020). Liu et al. (2014) suggested that the percentage of buildings, vegetation, and sky in an image is an effective landscape element affecting the perception of the soundscape. Verma et al. (2020) explored the relationship between visual features and perceptual attributes through Pearson correlation coefficients. Their results demonstrated the feasibility of predicting perception by visual features. This visual-based soundscape perception relies on people's knowledge and life experience. For example, a crowded street scene would evoke the sound of horns and people talking and a park may associate natural sounds such as animal chirps and water flowing. That is, humans might envision the acoustic environment through a visual scene and their experience without being there and hearing the sounds. Conveniently, as mapping services and volunteered geographic information have grown in popularity (Yan et al., 2020), a vast number of geotagged photographs spanning every corner of numerous cities around the world have been collected and made available (Anguelov et al., 2010; Zhang et al., 2018; Biljecki and Ito, 2021). Additionally, computer vision algorithms based on deep learning have made significant progress, garnering widespread attention and success in a variety of fields due to their outstanding automatic learning and representation capabilities for image features (Hinton et al., 2012; Cao et al., 2020; Chen et al., 2021; Liu and Biljecki, 2022).

We propose a new method for low-cost, large-scale and high-resolution prediction and evaluation of urban soundscapes — by using SVI data, essentially introducing a new use case of this growing urban data source. First, we extract four types of SVI features based on computer vision and deep learning models: low-level features, semantic segmentation, object detection, and scene classification. Second, fifteen soundscape indicators are constructed from four aspects: sound intensity, soundscape quality, sound source, and perceptual emotion. The soundscape indicators of a large number of SVIs were scored via a comprehensive international crowdsourcing effort with multiple responses per image to converge towards a consensus. Third, a machine learning model is used to predict the soundscape indicators of approximately half a million SVIs, and the high-resolution distribution of the city-level soundscape is obtained. Within this method, we seek to answer the following research questions: (1) How to sense soundscapes with different indicators at a city-level while achieving a high spatial resolution? (2) Is it suitable to employ crowdsourced labeled SVI as a new data source for assessing the soundscapes? (3) What is the relationship between the visual elements of the SVI and the

soundscape indicators? Our work contributes to understanding the distribution of soundscapes, revealing the relationship between the urban visual environment and soundscape, which is beneficial to the improvement of the urban acoustic environment. Unlike most other perception studies in the urban environment, we collect perception data from multiple sources with the information on where the study participant lives, allowing us to compare the differences in perception depending on whether the respondent resides in the city in focus or not, another contribution in the field.

2. Related work

2.1. Soundscape sensing

Soundscape sensing is a part of urban sensing, which can be viewed as a collective of technologies that perceive and acquire information about physical spaces and human activities in urban areas (Shi, 2021), such as safety (Song et al., 2020), vibrancy (Tu et al., 2020), and sustainability (Wu and Biljecki, 2021). The soundscape was introduced as an acoustic standard to interpret perceptions of sound by people in certain environments (ISO/DIS 12913-1, 2014; Korpilo et al., 2023). Rather than traditional acoustic research, which is focusing on the physical quantity of sound (e.g. intensity, frequency, and amplitude), soundscape studies tend to investigate both positive and negative effects of sound components from a human perspective (Schafer, 1993). Thus, various perceptual descriptors such as pleasantness and eventfulness (Axelsson et al., 2010; Jo and Jeon, 2020), calmness and relaxation (Davies et al., 2014; Sudarsono et al., 2017; Zhao et al., 2022), and other characteristics, have been widely used to describe perceived affective quality. Specifically, Axelsson et al. (2010) proposed a principal components model to define soundscape properties, extracting eight typical descriptors: pleasant, unpleasant, eventful, uneventful, exciting, monotonous, chaotic, and calm based on 116 attributes.

The sound components are the main independent variables that significantly contribute to different perception indicators. By interpreting properties in 25 videos into two components: urban environments and social environments, Axelsson (2015) found that the sound components are essential factors in terms of predicting perceptual descriptors and soundscape quality. Moreover, Jo and Jeon (2020) investigate the differences in soundscape quality assessment between visual environment and audio-visual environment among thirty participants. According to the comparison, not only certain kinds of sound sources such as human sounds and natural sounds, are significantly related to positive perceptual indicators as well as high quality of soundscape, but visual elements can also determine the initial perception of urban soundscape quality, which are also revealed by other studies

that non-auditory factors such as openness, density and visual properties of urban spaces have considerable importance in soundscape assessment (Yong Jeon et al., 2011; Hong and Jeon, 2015).

Recently, with the emergence of urban multi-source urban big data, more and more soundscape assessment studies apply multi-source data to evaluate the spatiotemporal patterns of soundscapes in both acoustic and non-acoustic aspects, providing comprehensive assessment and mapping at scale (Radicchi et al., 2016; Gasco et al., 2019; Zhang et al., 2019). As one of the examinations, by extracting the emotional layer of soundscape from social media data and combining them with the perceptual layer generate by a survey of sound walk in certain acoustic environments, Aiello et al. (2016) explored soundscape distribution in wider geographical coverage. In addition, Salem et al. (2018) proposed a location-dependent model to predict audio mapping from block-level to country-level according to joint feature representation generated by audio, ground-level, and overhead image appearance. This line of work indicates that multi-source information fusion provides potential scalable accessibility to understand the spatial structure and perceptual constructs of urban soundscape, as well as the opportunity to investigate the interrelationships among different urban attributes, such as acoustic indicators, visual features, and human activity information.

2.2. Street View Imagery in urban studies

SVI has created an opportunity to power urban studies across multiple scales because of its wide coverage and fine spatial sampling (Biljecki and Ito, 2021). Various studies have used it to explore urban information among multiple cities: quantifying urban greenery (Long and Liu, 2017; Wu et al., 2020; Hawes et al., 2022), assessing travel quality (Ito and Biljecki, 2021; Ning et al., 2022), extracting building features (Zhang et al., 2021b) and especially the measurement of the perceptual indicator (Naik et al., 2014; Dubey et al., 2016; Guan et al., 2022), supported by computer vision techniques. As one of the key deep learning models in computer vision, semantic segmentation is widely used for urban feature extraction, converting two-dimensional images into indexes based on convolutional networks, such as YOLO, SegNet, VGGNet, DeepLab, and so on. Such supervised models should be trained by certain datasets like Cityscapes (Cordts et al., 2016), which divides urban elements into 19 categories (e.g., road, car, vegetation, and sky) and therefore can automatically analyze the feature and appearance of images with high scalability. Moreover, other CV models such as Object Detection and Image Classification can also extract High-level features from the images efficiently (Verma et al., 2020). These three techniques are employed by Ito and Biljecki (2021) to conduct research on bikeability, in which 12014 images are collected from Singapore and Tokyo, extracting HLF as dependent variables to predict

the perception indicators.

Besides, SVI has also enabled examining visual features from approximating the pedestrian perspective. Urban perception studies often involve SVI surveys to investigate the subjective feelings of participants (Nagata et al., 2020) or the professional assessment from experts (Hanibuchi et al., 2019; Tang and Long, 2019). The result of surveys can generalize the images by means of perceptual labels and quantify them into urban attributes, such as the scores of safety, lively, beautiful, wealthy, and negative attitudes (Ordonez and Berg, 2014; Min et al., 2019; Yao et al., 2019). Through detecting the relationship between urban appearance and urban attributes by deep learning, visual features in SVI have multiplied opportunities for predicting non-visual indicators at a city scale, such as housing price (Arietta et al., 2014), street quality (Tang and Long, 2019), cases of infectious diseases (Nguyen et al., 2020) and community vitality (Wang and Vermeulen, 2021).

Similarly, informational attributes from different areas have been shown to have a direct and substantial effect on the perception of a soundscape. Studies focusing on the association between sound types and human perception by Axelsson et al. (2010) and Aiello et al. (2016) indicate that a location with more human activities (e.g. sounds from people talking or playing) wound tend to be more pleasant than a place dominated by technological sounds (e.g. sounds from vehicles, machines or construction). Also, Verma et al. (2020) used SVI to extract the visual features of streets in one part of Mumbai, and relate their audio aspects. Among the results, the study suggests that certain acoustic and visual characteristics are related to individual attitudes, indicating that there is a strong correlation between soundscape and human perception. However, a major research gap exists — there is a challenge to integrate the subjective indicator of soundscape with the strengths of SVI and apply it to large-scale urban information prediction through automated algorithms. Further gaps include understanding the multifaceted relationships between predictors of soundscapes and perceptual differences among different demographic groups.

3. Methodology

We present a large-scale and high-resolution urban soundscape sensing method using SVI. The presented method contains three steps (Figure 1): (1) constructing soundscape indicators, where fifteen instances are constructed from four aspects: sound intensity, soundscape quality, sound source, and perceptual emotion, and each indicator was labeled with a large-scale survey by an online survey; (2) extracting visual features of SVI at four levels: pixel-level feature, object-level feature, semantic-level feature, and scene-level feature, in SVI based on computer vision; (3) building a soundscape prediction model, where the SVI features and

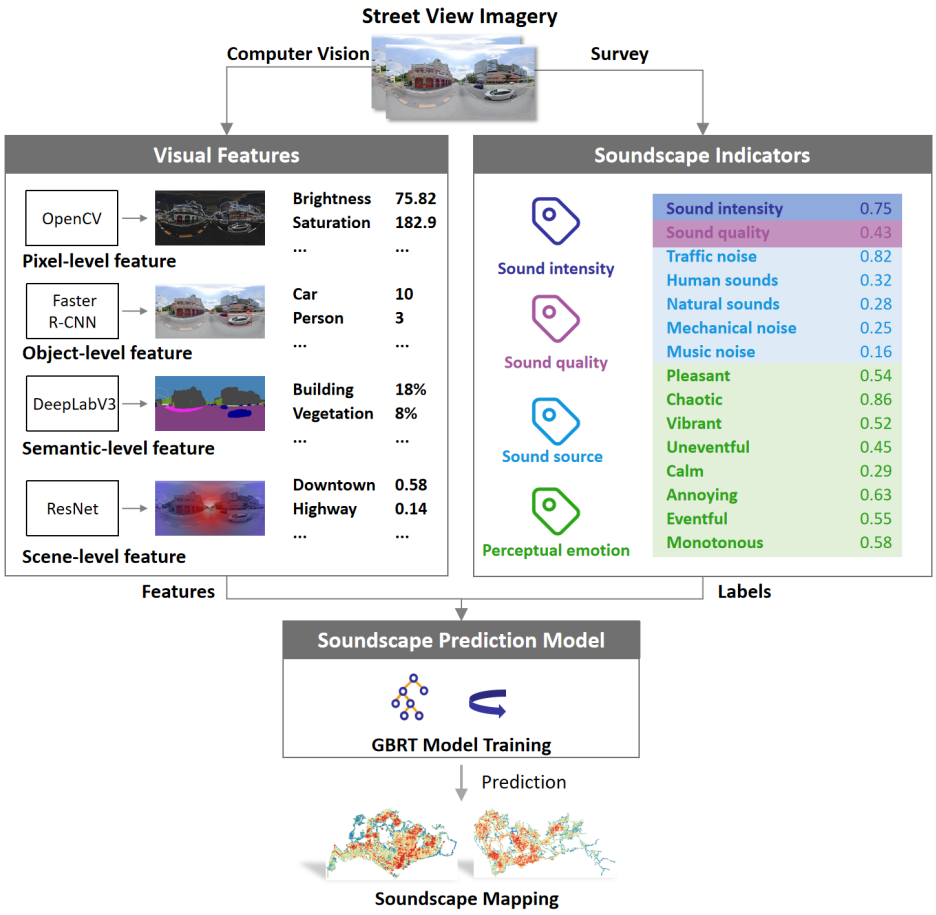

Figure 1: Overview of our workflow to sense soundscapes comprehensively from street view images. First, four levels of visual features are extracted using computer vision algorithms and deep learning models. Second, we constructed 15 soundscape indicators from four aspects and obtained the values of each indicator by scoring from volunteers. Third, a GBRT model is trained to predict the human soundscape perception of SVI in new urban areas.

soundscape labels are used as input to train the Gradient Boosted Regression Trees (GBRT) model, and by feeding city-scale SVI features into the trained model, the soundscape of a city can be mapped. In our work, we focus on two cities: Singapore and Shenzhen (elaborated in Section 4). These steps are detailed below.

3.1. Soundscape indicators

The soundscape is a conceptual framework for an acoustic or sound-related matter that involves both auditory components and human reactions (ISO/DIS

12913-1, 2014; Hasegawa and Lau, 2022). As illustrated in Figure 2, we construct a comprehensive soundscape indicator system from the acoustic environment to human reactions to evaluate urban soundscape, including fifteen indicators from sound intensity, sound scores, perceptual emotion, and sound quality aspects. It is worth noting that these indicators measure people's subjective perceptions, not the physical acoustic environment. Sound intensity is one of the most important indicators in the acoustic environment due to it being related to urban noise and it is the most intuitive one to people. Different built environments lead to different sound sources. According to the text describing the sound, Schafer (1993) comprehensively classifies the sound type, including seven major categories and the corresponding subcategories. Inspired by this research, we classify sound sources according to five subcategories: traffic noise, human sounds, natural sounds, mechanical noise, and music noise. We summarize human reflection into two aspects, one is people's overall evaluation of the sound, 'soundscape quality', and the second one is the perceptual emotion of different sounds. The perceptual emotion is summarised with eight subcategories: pleasant, chaotic, vibrant, uneventful, calm, annoying, eventful, and monotonous, which is obtained by principal component analysis of different sound perceptions (Axelsson et al., 2010). Finally, we established a soundscape index system of four categories and fifteen subcategories.

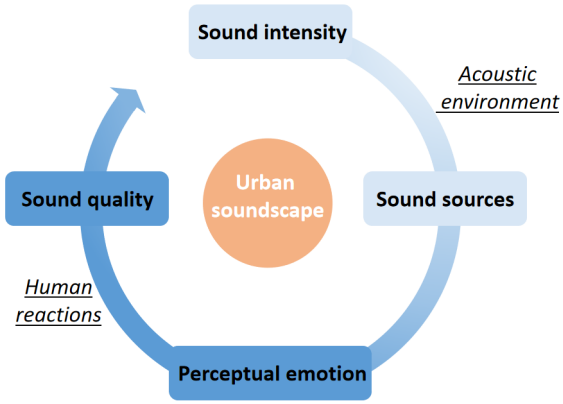

Figure 2: Urban soundscape indicator system from acoustics environment to human reaction.

A perceptual survey of the SVI is designed to collect the above-devised soundscape indicators. We classified participants into two types of groups: local and nonlocal for each city, which aims to investigate the impact of urban familiarity on the perceived results. We involved a total of 300 people to participate in the non-local group survey via Amazon Mechanical Turk, and 18 and 20 local residents from Singapore and Shenzhen, respectively, to participate in the local group

survey. The ethical aspects of this study have been reviewed and the survey was approved by the Institutional Review Board of the National University of Singapore (NUS-IRB-2021-906).

We selected 583 and 751 SVIs from Singapore and Shenzhen for the survey. The selection of the images was based on two principles: it should include as diverse scenes as possible and cover as many urban areas as possible based on the spatial location of the images. As highlighted in Figure 3, we combine the comparison and scoring method to obtain the value of each soundscape indicator. The range for each soundscape indicator value is from 1 to 5. The advantage of this method is that it can avoid data bias caused by subjectivity and randomness. The content of the soundscape perception survey is shown in Table 1.

The 15 soundscape indicator values for each SVI will be calculated based on independent and comparative fractions. The independent fractions for each scene and each indicator are derived from their average scores, which have been filtered to remove outliers. The comparative fractions are based on the 'win' and 'loss' scores of each image after being compared with other images. We define that when the indicator scores of the image are higher, equal, and lower than that of the image being compared, it would be scored 1, 0.5, and 0, respectively. Therefore, the definition of independent scores (I) and a comparative fraction (C) according to each perception indicator (a) and its score (Q) would be:

$$I_a = \frac{1}{5}\left(\frac{\sum_{t=1}^{T} Q_a^t}{T}\right) \tag{1}$$

$$C_a = \frac{1 * h_a + 0.5 * e_a + 0 * l_a}{T} \tag{2}$$

where T is the total number of times the image was been compared, h is the number of times the score of an image was higher than its paired image, while l is the number of times that an image was lower than its paired image, and e is the number of times when an image's score is equal to its paired image. Overall, the sum of h, l and e equals to T. Finally, we normalized the final perceptual indicator score (P) to between 0 and 1 of each image as:

$$P_a = \frac{1}{2}\left(I_a + C_a\right) \tag{3}$$

which will further be used as training indicators of each image for the soundscape perception model.

Figure 3: Snapshot of the first question in the survey (for one pair of scenes), emphasizing the comparison and scoring method.

3.2. Visual feature from street view imagery

Visual features of SVI are the important elements of the urban built environment, which include color, vegetation, architectural forms, urban scenes, and so on. The pixel-level, object-level, semantic-level, and scene-level visual features are constructed by using computer vision technologies. Specifically, the pixel-level features characterize the overall impression of SVI (e.g. brightness and saturation), which affects people's emotional perception. The object-level feature is to obtain the number of objects in a SVI, such as cars and people. The semantic-level feature is to obtain the proportion of pixels of different semantic items in a SVI, such as the proportion of vegetation, the number of vehicles, etc. The scene-level feature means the probability of scene semantics, such as parks or highways.

As shown in Table 2, the extraction of visual features is conducted through three types of pretrained deep learning models. The pixel-level feature was retrieved by the algorithms from the OpenCV library, including hue, saturation, lightness, and values edge detection features. For the task of the object-level feature extraction, Faster R-CNN (Ren et al., 2015), a model trained on COCO dataset (Lin et al., 2014), is used to identify and calculate the number of elements within 91 object types (e.g. person, bus, traffic light). The semantic-level feature extraction task relies on the DeepLabV3+ model (Chen et al., 2018) trained on Cityscape dataset

Table 1: The content of soundscape perception survey. For each question, participants are asked to express their preference on two five-point Likert scales for each soundscape indicator.

Question	indicator	Scale (from 1 to 5)
1. Overall, what is the general sound intensity (noisy or quiet) and sound quality (feeling good or bad) you feel from the two scenes above?	Sound intensity, sound quality	[very noisy, ..., very quiet], [feeling very bad, ... , feeling good]
2. For the following sounds types, to what extent do you presently feel them in the above two scenes	Traffic noise (cars, buses, trains, airplanes, etc.), human sounds (conversations, laughter, children playing, footsteps, etc.), natural sounds (birds, water, wind, etc.), mechanical noise (construction, industrial, machinery, etc.), music noise (bars, amplifiers, etc.)	[do not feel at all, ..., dominates completely]
3. For the following perceptions of the sound environment, to what extent do you agree or disagree these feelings are consistent with the two scenes above?	Pleasant, chaotic, vibrant, uneventful, calm, annoying, eventful, monotonous	[strongly disagree, ..., strongly agree]

Table 2: Summary of feature extraction models and algorithms

Feature	Model/Lib	Dataset	Features
Pixel-level features	OpenCV	–	Hue, Saturation, Lightness, Values Edge
Object-level features	Faster R-CNN	COCO	91 object types (person, bus, truck, etc.)
Semantic-level features	DeepLabV3	Cityscape	19 categories (road, vegetation, sky, etc.)
Scene-level features	ResNet	Places365	365 scene categories (highways, parks, downtown, etc.)

(Cordts et al., 2016), which includes more than 19 classes of labels (e.g., sky, vegetation, building, etc.) marked from ground level images. Lastly, in order to predict the probability of scene properties in a SVI, Places365 dataset (Zhou et al., 2017) is used to train the ResNet model (He et al., 2016) with 365 scene classes, such as highways, parks, downtown, etc. Based on SVI, explore the relationship between visual features and people's perceptions which is aimed at identifying key features in the visual feature which trigger particular perception (Herzog et al., 1976; Verma et al., 2020).

3.3. Soundscape prediction model

Predicting each soundscape indicator is regarded as a supervised regression task. Gradient Boosted Regression Trees (GBRT) is a machine learning approach that is based on a tree model, which has a solid performance in regression problems. Different from general tree models (e.g. decision trees), Gradient boosting combines weak 'learners' into a single strong learner in an iterative fashion (Friedman, 2002). This method provides an effective way of handling high-dimensional

features and can produce a reasonable prediction without hyper-parameter tuning.

The core of GBRT is that each calculation is to reduce the residual error of the previous one, to reduce these residuals, a new model can be built in the reduced gradient direction. In GBRT, each new model is built to reduce the previous residuals toward the gradient. The dataset is define by $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, $L(y, f(x))$ is the loss function. The number of leaves nodes in each regression tree is J, dividing its input space into J disjoint regions $R_{1m}, R_{2m}, \cdots, R_{jm}$, b_{jm} was estimated for each region. The regression tree $g_m(x)$ is expressed by the following:

$$g_m(x) = \sum_{j=1}^{J} \left(b_{jm} I\right), x \in R_{jm}, I\left(x \in R_{jm}\right) = \begin{cases} 1, & x \in R_{jm} \\ 0, & \text{else} \end{cases} \tag{4}$$

The prediction accuracy of the GBRT model is mainly affected by the number of regression trees(M) and the learning rate. In general, as the M increases, the model's prediction accuracy would improve, however, too many trees may result in additional computation and overfitting. When the number of regression trees is fixed, increasing the learning rate may improve the prediction accuracy faster, but a lower learning rate can achieve better prediction accuracy. The training data consists of SVI visual features as input variables and corresponding soundscape indicators as target values. We used a total of 482 SVI visual features as input to predict 15 distinct soundscape indicators.

4. Study area and dataset

4.1. Study area

Singapore and Shenzhen (Figure 4) were selected as research areas. Singapore covers 724 km^2, 540 of which are built-up areas. It includes 5 administrative districts: East region, North-east region, North region, Central region, and West region. Each region in Singapore has a high level of development intensity, and traffic noise is the main source of the noise. It should be noted that Singapore Changi Airport is located in the East region and may be subject to airplane noise. The facilities of the Port of Singapore, the world's second-busiest port, are mostly in the Central region and West region. The industrial center is mainly in the West region. There are many construction sites here as a result of urban development, which could generate a lot of noise.

Shenzhen covers 1995 km^2, 800 of which are built-up areas It includes 10 administrative districts: Luohu, Futian, Nanshan, Yantian, Baoan, Longgang, Guangming, Longhua, Pingshan, and Dapeng. Futian and Luohu are recognized as the city center, and they contain high buildings and are characterized by dense employment, the main source of noise here is traffic. Nanshan is a high technology district

with many innovative companies and factories. Shenzhen International Airport is located in Baoan, where the noise is driven by the airport. Shenzhen port is located in Nanshan and Yantian. There are many trucks near the port, which is a significant source of traffic noise.

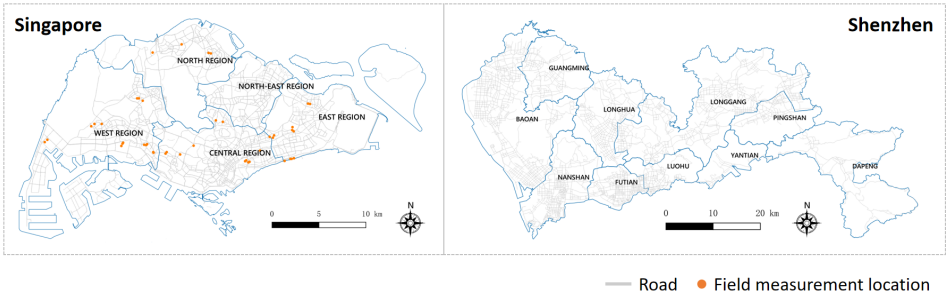

Figure 4: Study area: two major cities with diverse soundscapes. The figure also includes the locations of the sound measurements. Map data: (c) OpenStreetMap contributors.

4.2. Street View Imagery

This study assesses urban soundscapes using SVI data from Baidu and Google, the two major sources of such data. Among these, Singapore data comes from Google Street View[1], while Shenzhen data comes from the street view service of Baidu Maps[2]. We generate sample points every 50 meters within the study area and search for the nearest panoramic SVI. Finally, we obtained 270,055 and 329,802 panoramic SVIs in Singapore and Shenzhen, respectively. These images are used for both the survey and computer vision portion of the method.

4.3. Field audio data

Audio data were collected from 43 randomly selected investigation points in Singapore (Figure 4). Three-minute video clips, 4-10 SVIs, and a three-minute recording of variations in sound intensity for each investigation point are collected. The devices used in the collection include a Sound Level Meter ($UT353BT$) for sound intensity recording and a smartphone for the shooting of videos and SVI. We have released the collected data openly. The collected data can be obtained at https://github.com/ualsg/Visual-soundscapes.

[1]https://www.google.com/maps/
[2]https://map.baidu.com/

Table 3: Sound intensity prediction accuracy in different models.

Modle	Singapore		Shenzhen	
	MAPE (%) ↓	R^2 ↑	MAPE (%) ↓	R^2 ↑
DTR	30.1187	0.3049	29.8720	0.4321
KNR	30.8807	0.4838	24.3095	0.6237
SVR	26.8931	0.5650	23.8692	0.6596
BR	24.7251	0.5924	23.2427	0.6731
RR	28.0802	0.5973	23.6873	0.6804
RF	24.0810	0.6437	22.4218	0.6889
GBRT	**21.5036**	**0.6808**	**21.8282**	**0.6936**

5. Results and analysis

5.1. Soundscape prediction result

5.1.1. Model comparison

To demonstrate the superiority of the GBRT model, we developed the model by comparing different machine learning models of Decision Tree Regression (DTR), K-Neighbors Regression (KNR), Ridge Regression (RR), Support Vector Regression (SVR), Bagging Regression (BR), Random Forest Regression (RF) and GBRT. The dataset was constructed by the SVI used in the surveys covering both cities, with 80% of the SVIs serving as the training dataset and 20% serving as the test dataset. The models are validated using K-fold cross-validation, which breaks the data into K folds, and each fold is used as a test set. K=10 is used in this study. The mean absolute percentage error (MAPE) and coefficient of determination (R^2) were used to assess the model.

Taking sound intensity prediction as an example, we calculated the average value of 10 experiments for each metric. The findings are summarized in Table 3. Overall, MAPE in Singapore and Shenzhen are between 21.50 and 30.12, R^2 is between 0.30 and 0.69, and R^2 in Shenzhen is higher than in Singapore. The DTR model performed the worst in both datasets. The MAPE and R^2 of the GBRT model have the best performance in the Singapore and Shenzhen dataset. As a result, GBRT is chosen as the prediction model.

5.1.2. Prediction result evaluation

The MAPE and R^2 are used to evaluate the prediction results of the 15 soundscape indicators predicted by the GBRT model. The MAPE is shown in Figure

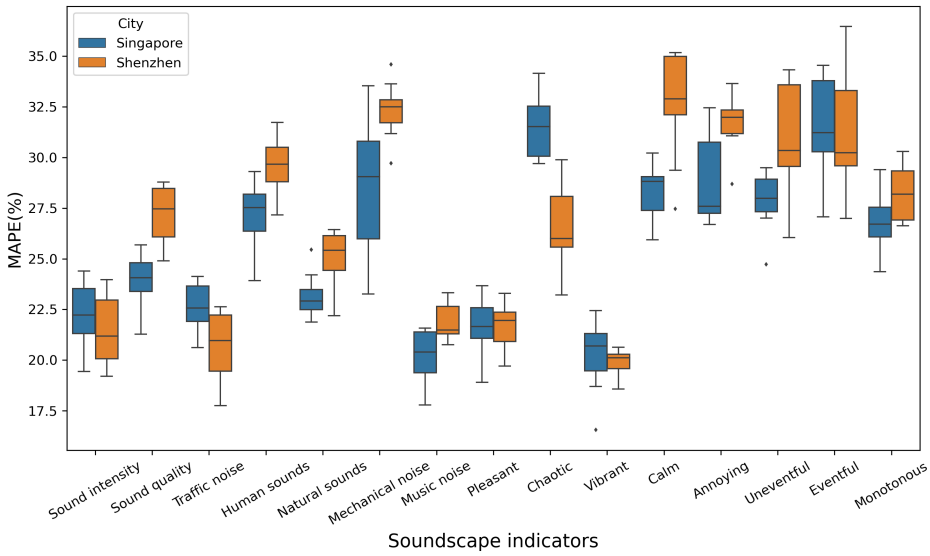

Figure 5: The MAPE of soundscape indicators prediction model for the two cities.

5. Overall, the various MAPE of soundscape indicators is quite different. Taking Singapore as an example, the median MAPE of sound intensity is 22.22, while the median MAPE of chaotic reaches 31.52. There could be two reasons for this difference. First, it is related to the distribution of the values of the soundscape indicators. For example, the musical noise indicator value is typically tiny in most situations, and the same absolute inaccuracy may result in a higher MAPE. Second, the predictability of various indicators of soundscapes is different. In addition, the MAPE of soundscape indicators in different cities are different. In Shenzhen, the median of MAPE of chaotic is 25.99, whereas in Singapore, the value is 31.52, probably due to the different distribution of features in the different datasets. For example, Shenzhen has a high proportion of chaotic scenes, and the features are more obvious, implying that this indicator has high predictive accuracy.

The R^2 of the prediction model is shown in Figure 6. Similar to MAPE, the R^2 varies according to the soundscape indicators. Taking Shenzhen as an example, the R^2 of traffic noise and sound intensity are higher with the median R^2 0.74 and 0.69, respectively, while music noise and monotonous have lower R^2 with medians of 0.20 and 0.18 respectively. This result demonstrates that humans have a greater sensitivity to the perception of sound intensity, traffic noise, chaotic, while the perception of certain attributes such as music noise, uneventful, and monotonous, is diminished, which is consistent with our expectations and previous study(Axelsson

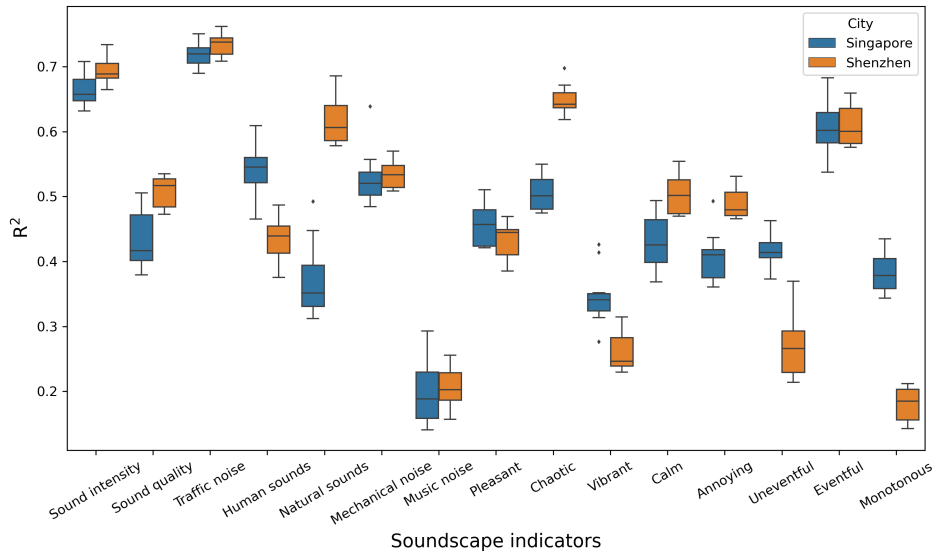

Figure 6: The R^2 of soundscape indicators prediction model for the two cities.

et al., 2010). Moreover, there is another possibility that the lower R^2 value is due to the fact that individuals with varying knowledge of the soundscape indicators have conflicting opinions, resulting in the R^2 value for the same soundscape indicator varies between cities. For example, the R^2 of nature sounds is 0.61 and 0.35 in Shenzhen and Singapore, respectively, which is probably because the Shenzhen dataset contains more components that significantly reflect natural sounds.

5.1.3. Prediction result validation

The predicted soundscape indicators are the acoustic environment that people perceive from the SVI. To verify the sensibility of using SVI to predict the soundscape, we calculated the correlation between the predicted perceived sound intensity and the field measurements. As mentioned in Section 3.2, the visual features of pixel-level, object-level, semantic-level, and scene-level are extracted from the field collected SVI. Pixel-level, object-level, and semantic-level features are extracted from panoramic SVIs, while scene-level features are extracted from several SVIs taken from various angles, and their average values are used as the final results. These visual features are fed into the trained model to obtain the predicted sound intensity. The ground-true sound intensity is the average of the 3-minute sound intensity recordings collected in the field after removing outliers. The predicted and measured sound intensity correlations are shown in Figure 7, with 0.48 as the R^2 of the Singapore dataset, indicating the reliability of using SVI to assess

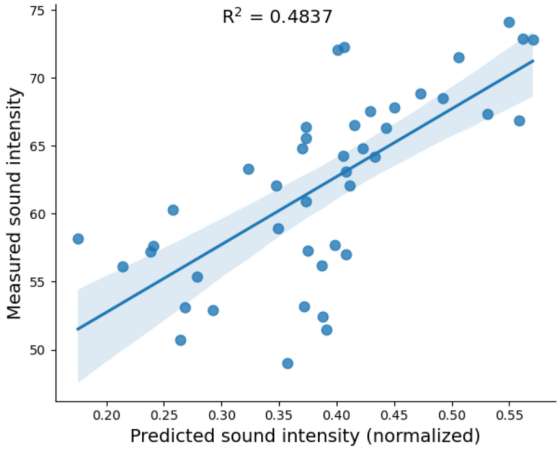

Figure 7: Relationship between predicted and measured sound intensity.

soundscapes. Besides, there are some field measurements of sound intensity that differ significantly from predicted values for two reasons. The SVI data collected in the field and the data provided by Google/Baidu have different shooting angles, resulting in different components of visual features. In addition, the intensity of sound in one place may vary over time and the intensity of sound measured in three minutes is not necessarily representative.

5.2. Soundscape mapping and spatial analysis

All SVI features of the two cities are ingested into the trained model to obtain the soundscape mapping, which can facilitate the investigation of soundscape spatial distribution. The spatial unit is a hexagonal grid divided by a geospatial indexing system H3[3], and the resolution is level 9, with an average hexagon area of 0.105 km^2. Each indicator value for a hexagon is the average value for all SVIs within a certain hexagonal grid.

5.2.1. Sound intensity mapping

Sound intensity is the most concerned and sensitive soundscape indicator by residents. Therefore, this indicator is chosen for further analysis. The sound intensity distribution in Singapore is shown in Figure 8. The red units refer to areas with higher sound intensity, mostly concentrated in the suburbs, used for infrastructure construction (e.g. Tuas, ①). Specifically, some units around a highway, such as

[3]https://h3geo.org/

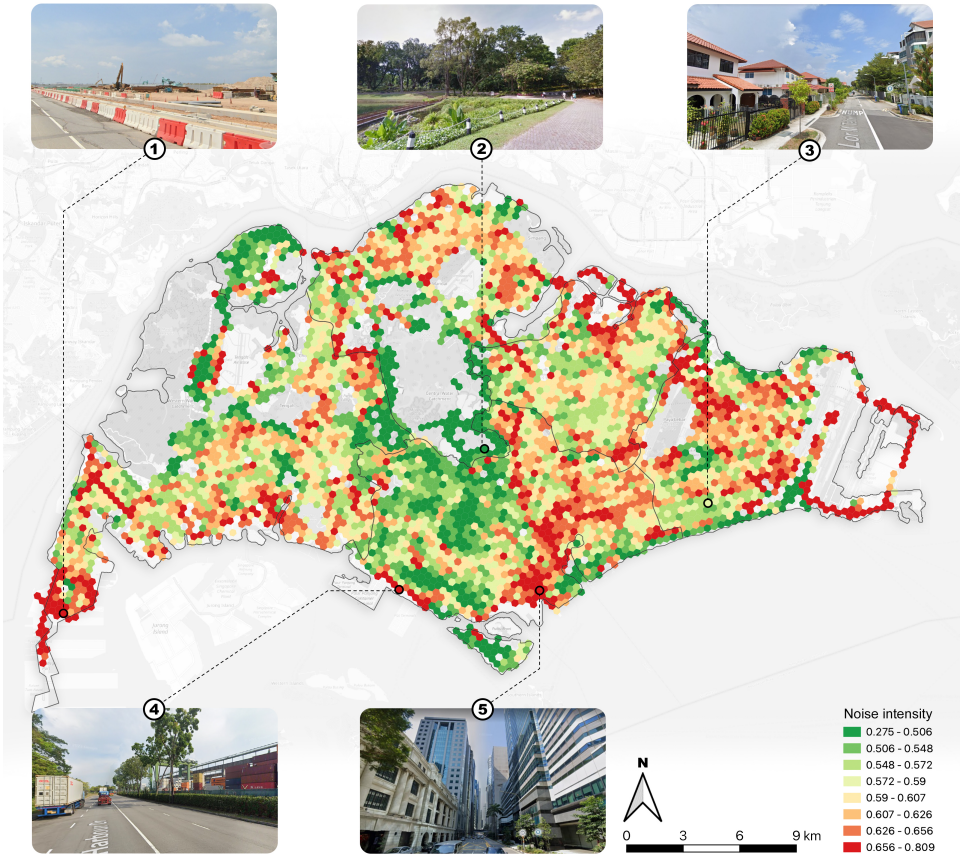

Figure 8: Sound intensity distribution in Singapore. Map and image data: (c) OpenStreetMap contributors, Google Street View.

the West Coast Highway next to the port (④), serving for heavy logistics traffic, have a significant impact on noise intensity. In addition to this, the bustling core business district (⑤), indicates a sound intensity at a high level. Low-intensity areas are identified in parks with more vegetation, as well as tourist destinations such as the Singapore Botanic Gardens (②), East Coast Park, etc. Typical residential neighborhood areas such as Bedok (③) have median sound intensity levels. In general, industrial areas and central business districts have a high noise level, but tourist attractions and residential areas have a low noise level. The distribution of sound intensity is highly correlated with urban function, in line with Chew and Wu (2016), who exposed the levels of noise differ from various land use.

Figure 9 depicts the spatial distribution of sound intensity in Shenzhen. In gen-

eral, the distribution of sound intensity is higher in the north and lower in the south. Most of the high-intensity areas are concentrated along highways, whereas the low-intensity zones are mostly concentrated along parks and the coastline, which is consistent with our expectations. Specifically, the areas with high levels of sound intensity include construction sites and highways, such as Qianhai (marked ①) in Nanshan District, a new development zone, and the Baoan expressway (marked ②). Typical low-sound-intensity areas, such as the Wutong Mountain Park and Dapeng (marked ⑤), are heavily vegetated tourist destinations. Longhua has a large number of residential neighborhoods (marked ③) with median sound intensity. To our surprise, the most prosperous districts, Nanshan, Futian, and Luohu, have lower sound intensity than expected. This could be because these areas are also well vegetated, as shown in the corresponding SVI (marked ④), which could have softened people's perception of sound intensity. This is consistent with the findings of Van Renterghem (2019), who proposed that vegetation can strongly improve environmental noise perception.

5.2.2. Typical soundscape indicators mapping

The soundscape quality, natural sounds, traffic noise, pleasant, and annoying were chosen as typical soundscape indicators, and the results are displayed in Figure 10 as soundscape maps, a key result of this work. For Singapore, the areas with better soundscape quality (green) are mainly distributed near parks, such as Sentosa, Reservoir Park, and East Coast Park. Areas with poor soundscape quality are mainly found in central business districts, industrial areas, and suburbs with more construction sites. Natural sound values are generally greater in areas adjacent to parks with more vegetation. Interestingly, although the soundscape quality of the industrial area is lower, it is also higher for natural sound indicators due to the low building density, such as in the Tuas area. The distribution of traffic noise and annoying was similar, with higher values concentrated near highways, industrial areas, and central business districts. The spatial distribution of pleasant is similar to natural sounds, but the difference is this indicator is also higher in residential environments.

For Shenzhen, spatial heterogeneity in the distribution of soundscape quality is significant, and areas with good soundscape quality are primarily concentrated around parks with vegetation cover, such as Dapeng and Wutong Mountain Park in Luohu District. However, the poor soundscape quality is concentrated in the vicinity of the highways. To our surprise, pleasant is higher in the three developed urban areas of Nanshan, Futian, and Luohu, although it is prosperous with more traffic noise. This result is because the developed urban area in Shenzhen have more greenery and are more orderly, providing a more pleasant environment for residents.

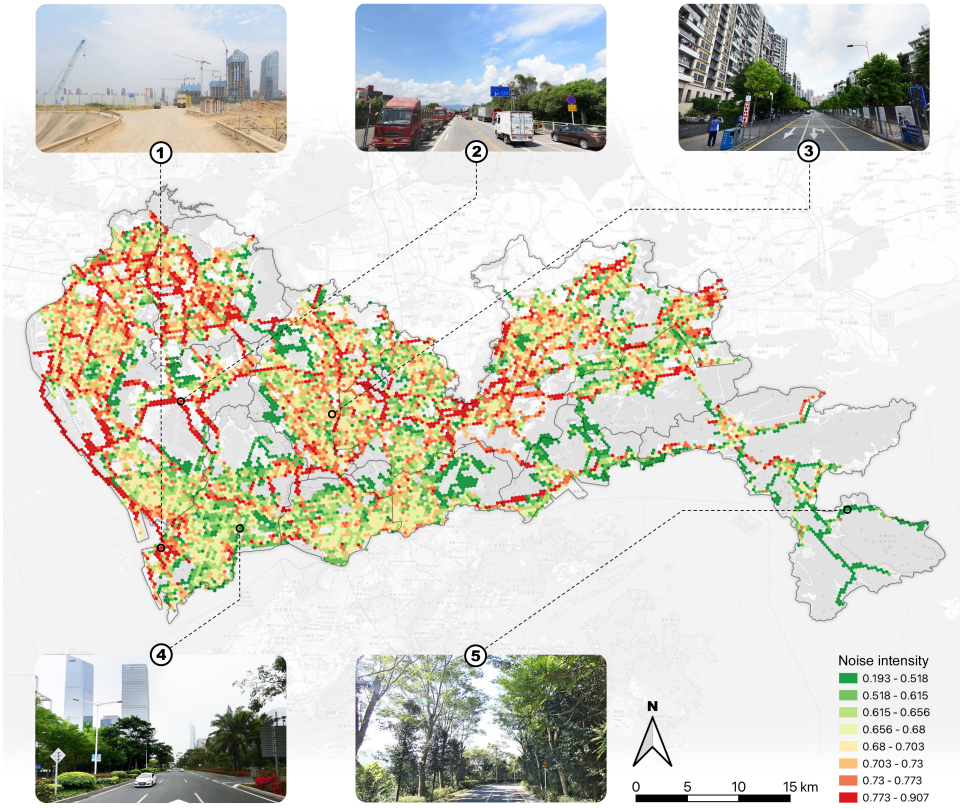

Figure 9: Sound intensity distribution in Shenzhen. Map and image data: (c) OpenStreetMap contributors, Baidu Maps.

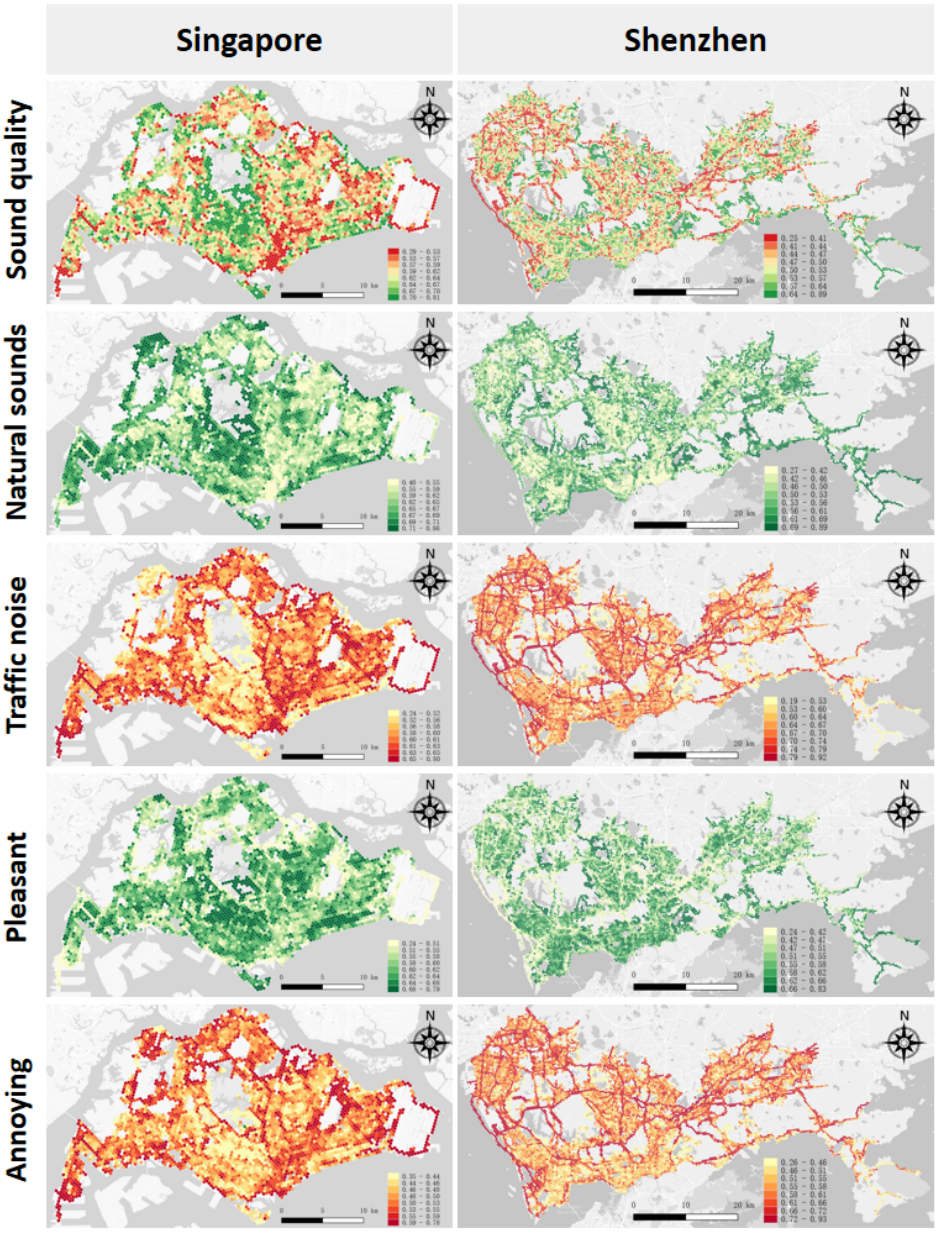

Figure 10: Spatial distribution of typical soundscape indicators. Map data: (c) OpenStreetMap contributors.

5.3. Relationship between visual features and soundscape

A multiple regression model is used to investigate the contribution of visual features to soundscape indicators. To improve the model's interpretability, we integrated and filtered the 482 visual features into 28 predictor variables (Table 4). Variable selection is critical for identifying the optimal subset of predictors and minimizing redundancy and noise. For each of these models, the Backward Elimination approach is used for variable selection (Derksen and Keselman, 1992). The Backward Elimination procedure includes: (a) selection of the significance level (in this study, a 95% confidence interval or 0.05 is considered); (b) fit the model with all the features selected; (c) remove the variable with the largest p-value; (d) fit the model with the remaining variables; and (e) repeat steps c and d until all the variables are less than the selected significance level or the number of variables is less than 10.

Table 4: Regression analysis variables

Features	Variable	Definitions
Pixel-level	canny_edges	The ratio of pixels detected as edges to the total number of pixels in the SVI
	hue_mean	The mean value of the hue dimension in the SVI
	hue_std	The standard deviation of hue dimension in the SVI
	saturation_mean	The mean value of the saturation dimension in the SVI
	saturation_std	The standard deviation of saturation dimension in the SVI
	lightness_mean	The mean value of the lightness dimension in the SVI
	lightness_std	The standard deviation of lightness dimension in the SVI
Objest-level	car_object	Total number of cars in the SVI
	bus_object	Total number of buses in the SVI
	person_object	Total number of person in the SVI
	truck_object	Total number of trucks in the SVI
	motorcycle_object	Total number of motorcycles in the SVI
	other_object	Total number of other remaining objects in the COCO dataset in the SVI
Scene-level	neighborhood_scene	Probability of the image being classified as neighborhood
	highway_scene	Probability of the image being classified as highway
	parking_scene	Probability of the image being classified as parking
	downtown_scene	Probability of the image being classified as downtown
	construction_site_scene	Probability of the image being classified as construction site
	industrial_area_scene	Probability of the image being classified as industrial area
	park_scene	Probability of the image being classified as park
	street_scene	Probability of the image being classified as street
	field_wild_scene	Probability of the image being classified as field/wild and forest road
	other_scene	Probability of the image being classified as other remaining classes in the Place365 dataset
Semantic-level	sky_semantic	Percentage of pixels classified as sky
	nature_semantic	Percentage of pixels classified as vegetation
	human_semantic	Percentage of pixels classified as person and rider
	vehicle_semantic	Percentage of pixels classified as car, truck, bus, train, motorcycle, and bicycle
	building_semantic	Percentage of pixels classified as building, wall, and fence
	other_semantic	Percentage of pixels classified as other remaining classes in the Cityscapes dataset

The results of the multivariate regression analysis between visual features

and soundscape indicators are shown in Figure 11. We selected the top 6 SVI features including positive contribution (red bar) or negative (green bar) to rank and list. The length of the bar indicates the value of the betas coefficients and the '*' implies the significance level. In general, SVI features contribute variably to different soundscape indicators. For the sound intensity, *Building_semantic* and *sky_semantic* play the most significant positive correlations, however, *nature_semantic* and $field\_wild\_scene$ are the strongest negatively correlated visual features. The result demonstrates *nature_semantic* is positively correlated with the sound quality score, while $construction\_site\_scene$ and $truck\_object$ are negatively correlated with soundscape quality, which is consistent with our common sense and expectations. In addition, two pixel-level features, $lightness\_mean$, and $canny\_edges$, occur in the sound quality list, demonstrating that these intuitive impressions of SVI, such as lightness, can have a significant impact on how individuals perceive sound quality.

In terms of sound sources, traffic noise and mechanical noise are positively influenced by similar visual features, such as *building_semantic* and *sky_semantic*. The mechanical noise is related to human-related visual features($person\_object$, $neighborhood\_scene$), due to mechanical noise being generally made by humans. The visual elements with the strongest positive and negative correlations to natural sounds are *nature_semantic* and *downtown_scene*, respectively, which is consistent with our expectations. Human sounds and musical noise have strong positive correlations with $person\_object$, *nature_semantic*, and *building_semantic*, while negatively correlated with $field\_wild\_scene$ and *highway_scene*, due to both of these sound sources are related to the distribution of the crowd. It is worth noting that the assessment of sound sources is mainly based on human experience rather than directly seen objects, which might lead to some perception bias (Zhang et al., 2021a). For instance, even if there are no vehicles on the highway, people's experience will lead to the perception that such a situation entails a significant level of traffic noise.

Regarding perceptual emotion, we observed that pleasant is significantly positively correlated with *nature_semantic* and *building_semantic*, and negatively correlated with *highways_scene*. This result validates the finding by Hong and Jeon (2017) that a pleasant perception of natural sounds has a positive effect and is negatively associated with vehicle sounds. Chaotic and eventful are positively affected by the same visual features, e.g. $person\_object$ and $neighborhood\_scene$. Vibrant and clam have a positive relationship with *nature_semantic*, interestingly, the car_object feature, is positive for Vibrant but negative for clam. The uneventful and annoying scene shows significant associations with mostly similar visual features, for example, the *nature_semantic* and *building_semantic* have a positive influence, while the *downtown_scene* have a negative influence on these sound-

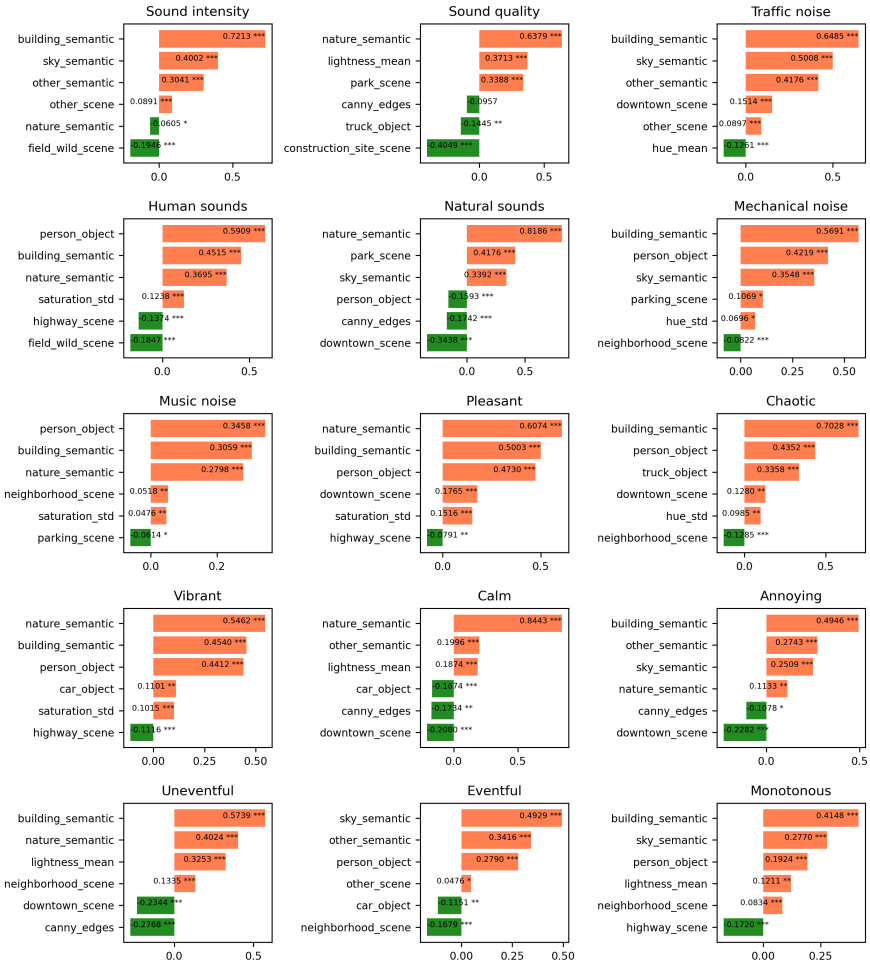

Figure 11: The results of the multivariate regression analysis between the visual features and soundscape indicators. (***p <0.001, **p <0.01, *p <0.05).

scape indicators.

5.4. Correlation of soundscape indicators

To explore the relationship between different soundscape indicators from the SVI survey, we calculated the cross-correlation matrix, as illustrated in Figure 12. The four types of soundscape indicators have been marked: I (sound intensity), Q (sound quality), S (sound source), and P (perceptual emotion). Overall, there is a strong positive correlation among sound intensity, traffic noise, chaotic, mechanical noise, eventful, and annoying. These soundscape indicators are all noise-related aspects that will elicit negative emotional reactions from participants. There is also a strong positive correlation among human sounds, music noise, vibrant, and pleasant, which represent people-related soundscape indicators. Moreover, a positive connection among natural sounds, sound quality, and calm, these soundscape indicators are high quality sound-related. On the contrary, there is a strong negative correlation between noise-related indicators (i.e. sound intensity, traffic noise, chaotic, mechanical noise) and high quality sound-related (i.e. natural sounds, sound quality, and calm). Specifically, the result indicates a significant positive association between sound intensity and traffic noise (r=0.73), chaotic (r=0.71), mechanical noise (r=0.68), and eventful (r=0.57), while a negative correlation is shown between sound intensity and natural sound (r=-0.44) and clam (r=-0.55). On the other hand, sound quality shows a contrary relationship with other attributions. There is a significant positive correlation between sound quality and natural sounds (r=0.51), clam (r=0.53), while a negative relationship between noise quality and traffic noise (r=-0.41), mechanical (r=-0.43), and chaotic (r=-0.47), respectively. For the relationship between sound source and perceptual emotion. Just as we expected, natural sounds have a positive correlation with calm (r=0.57), while chaotic (r=-0.49), which is similar to the result obtained by Verma et al. (2020). For the correlations between the perceptual emotion, chaotic and eventful were positively and significantly associated with each other, while they are negatively correlated with pleasant, vibrant, and calm. This finding is consistent with Aiello et al. (2016), which evaluated soundscapes using social media data.

6. Discussion

6.1. Application of soundscape sensing

We propose large-scale high-resolution soundscape perception relying on SVI data, which makes it possible to observe city-scale soundscapes from a macroscopic perspective. There are many potential applications based on the results and our method. For residents, the soundscape map can provide home buyers with a reference to find areas away from the noise and with high-quality soundscapes. For urban planners, the soundscape distribution allows urban planners to optimize the

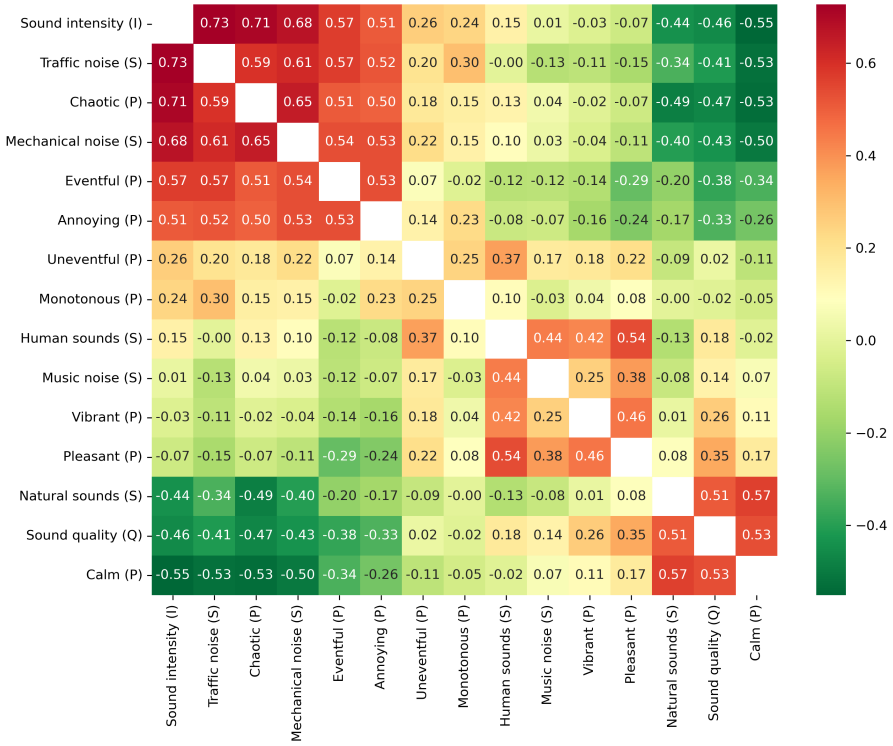

Figure 12: Cross-correlation between the soundscape indicators.

acoustic environment for various areas. Furthermore, in Section 5.3, the analysis result of the relationship between visual features and soundscape indicators, such as the contribution of vegetation, and building to specific soundscape indicators, would support urban design theory and practice. For theoretical implications, the soundscape is an important part of the physical properties of cities, and soundscape enriches place semantics, which will help researchers understand underlying urban heterogeneity patterns and reveal the impact of urban functions. In terms of technology, the potential application is that our results can inspire the generation of urban soundscapes directly (and solely) using SVI, such as generative adversarial network(GAN) (Wu and Biljecki, 2022). It can generate soundscapes according to the designed urban scenes, evaluate urban design schemes in visual and acoustic aspects, and improve the efficiency of urban design.

6.2. Soundscape perception bias

The term 'perception bias' refers to the disparity between the indicators predicted by a model and our real measurement or common sense. Some previous

research using SVI for urban study have mentioned perception bias, such as urban crime (Zhang et al., 2021a), playability (Kruse et al., 2021), and built environment (Wang et al., 2019). The majority of studies consider perception bias to be a study limitation, as it may induce model errors. However, we consider that exploring and understanding the bias may shed some insight on urban planning. As mentioned in Section 5.2, the actual sound intensity in Shenzhen's downtown region is quite high, due to dense crowds and heavy traffic flow. However, due to the high roadside vegetation coverage and good landscape quality in these areas, people's perception of sound intensity is reduced, and the pleasant and sound quality scores are also high. In other words, vegetation and better landscape quality reduce people's perception of noise. Understanding these perception biases is helpful for people to understand the relationship between soundscape and visual elements, which can guide urban landscape design and improve the quality of urban soundscape.

Although the differences in age of training participants mentioned by Wang et al. (2019) may not lead to bias, to examine perception bias in people with different living cities, we compared the local group to the non-local group, which was described in Section 3.1. Taking Shenzhen as an example, we select scenes from the Shenzhen dataset with reflect high levels of noise, such as highways and downtown. Both local and non-local groups have been invited to score these SVIs independently. Each image is evaluated in the same manner, and the result is shown in Figure 13. While there is little difference in soundscape perception between locals and non-locals on average, locals' scores are more consistent due to their shared perceptions of the city. For example, while both locals and non-locals have medians of '4', their first and third quartile are '4' and '5', respectively, but non-locals have two quartile of '3' and '4', with lower adjacent values of '2'. Additionally, locals score more precisely, and more specific soundscape indicators are more likely to be classified as maximum '5' or minimum '1', whereas non-locals may have more moderate values '3', as the indicator chaotic demonstrates. In addition, the soundscape is the perception of people based on their experiences, which allows for different insights. Overall, the perception bias due to differences in participants' backgrounds is acceptable. We hope that this detail will contribute to the body of knowledge of studying perception in urban informatics, as demographics are rarely accounted for.

6.3. Advantages of Street View Imagery

To enable large-scale and low-cost soundscape evaluation, several new data sources have lately been developed for soundscape evaluation, including social media data, complaint data, and 3D city models (Aiello et al., 2016; Tong and Kang, 2021; Stoter et al., 2020). As a wide-coverage, highly accessible data source, SVI

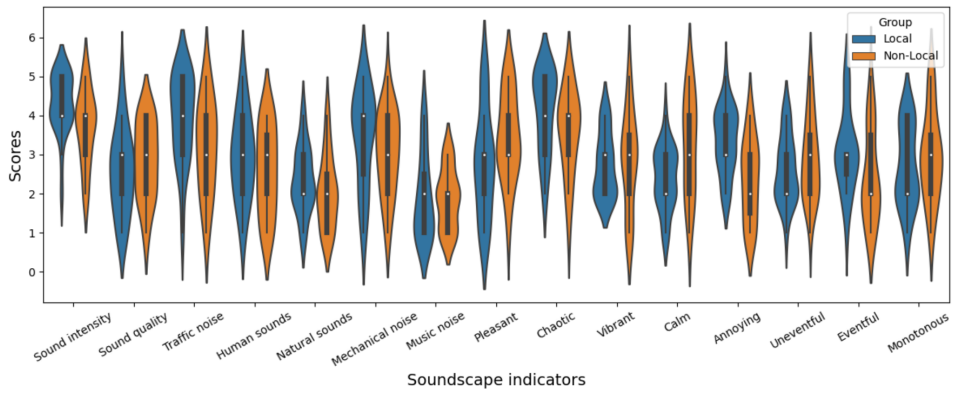

Figure 13: Perception differences between locals and non-locals.

data offers significant advantages for evaluating soundscapes. Its advantages include the following: (1) Precise geographic coordinate information. Commercial map service providers, such as Google and Baidu, collected data with very precise coordinate information, which can be accurate to the decimeter level (Anguelov et al., 2010). However, for the social media data with geotagged images or text, the coordinate information may be offset or even wrong, depending on the location of the information publisher (Fan et al., 2020). This shortcoming may introduce significant mistakes in the investigation of the spatial distribution of noise. (2) Wide-coverage and frequent-updates. Along with the commercial map service providers providing high-quality SVI data, crowdsourced SVI platforms (e.g. Mapillary, KartaView) have recently become popular. For places where commercial map service providers may not have data coverage, such as indoors, or in remote parks, crowdsourced platforms can supplement the data. While the data quality of crowdsourced SVI data is difficult to control (Mahabir et al., 2020), in comparison to other types of data, the data's availability, coverage, and update rate provide significant benefits. (3) Visual and auditory perception are inextricably linked. Several recent studies have exploited the relationship between sound and visual appearance for tasks such as sound generation (Salem et al., 2018; Aytar et al., 2016). These studies highlight the advantages of evaluating soundscapes in combination with visual elements. Therefore, evaluating the soundscape with visual data (e.g. SVI) can lead to higher interpretability.

6.4. Limitations and future work

We notice a few limitations of this work, which may be tackled in future efforts. First, although the urban soundscape of a particular area is generally consistent, such as the central business district being noisier than a typical park, the urban

soundscape is also dynamic, since the soundscape of a place varies greatly during the day (and night). However, the urban soundscape perceived by SVI data only describes a specific moment in time when the data was collected. In addition, with the development and renewal of the city, the soundscape environment is constantly changing, therefore, the evaluation results have a certain lag. Second, our work is a preliminary study of soundscape perception directly with non-acoustic data, we only use visual features to perceive soundscape. In fact, the soundscape of a place cannot be generalized from visual features alone. As Zhang et al. (2019) mentioned, a single big data for urban perception have certain biased. Future research can use multi-source big data, such as points of interest, and human mobility data, to evaluate the urban soundscape. Third, our SVI data come from different sources (e.g. Baidu, Google). Although some researchers assert that the results of analysis using local mapping services (e.g. Baidu Maps in China) can be replicated using GSV (Cheng et al., 2017), the predicted urban soundscape perception results may have some biases, mainly caused by two aspects. On the one hand, the different devices used to capture the SVI result in different aspect ratios, saturation, and brightness, etc., which can lead to perceptual bias. On the other hand, the different data distribution leads to the bias of urban soundscape maps. For example, Baidu Map's SVI is mainly distributed on main roads, with less coverage on some park paths, while GSV has more comprehensive coverage, which will lead to stronger traffic noise on Baidu Map than GSV. Therefore, it is not reasonable to compare and rank the soundscape quality of the two cities through SVI data.

7. Conclusions

We presented a new approach to understanding multi-dimensional soundscapes from street view imagery, a growing form of urban big data that has permeated through urban informatics but one that has not been used for such a purpose yet. Taking Singapore and Shenzhen as diverse examples, visual features were extracted from about half a million SVIs using a computer vision model based on deep learning. We have established fifteen soundscape indicators to comprehensively evaluate soundscapes, 1334 SVIs were evaluated using crowdsourcing, generating fifteen soundscape indicator labels. A machine learning model, GBRT, was developed to predict urban soundscapes and analyze spatial distribution. Additionally, we measured the actual sound intensity at dozens of locations to validate the model's reliability. We release this dataset openly to spur further efforts and future studies. To investigate the relationship between visual features and soundscape indicators, we developed regression models.

The result has shown that it is possible to predict and interpret the soundscape from SVI with machine learning at a reasonable accuracy. Additionally, we find

different visual elements have varying effects on predicting certain indicators, for example, vegetation tends to reduce the perception of sound intensity and invokes a pleasant feeling. There are combinations of high associations between various sound sources and perception elements of sound, such as traffic and machinery sounds increase chaotic and annoying impressions. This study brings the following contributions:

1. Our work elaborates how SVI data, a newly available data source, could be used directly for soundscape prediction and evaluation on a wide-scale, and yet at high-resolution and low-cost. Thus, the paper essentially introduces a new use case of SVI.
2. We demonstrate that crowdsourced labeled SVI can be used for soundscape prediction, and compare the contrast in perception among people with various backgrounds, a rarity in perception studies in urban informatics and related domains.
3. The relationship between urban sound sources and emotional perception, which are both visual, is examined.
4. We created high-coverage and high-resolution soundscape maps of Singapore and Shenzhen using solely SVI data, potentially paving the way for generating noise maps in a straightforward manner or supplementing existing noise maps by adding the qualitative aspect.

The benefits of this approach are multi-fold: the approach bypasses tedious ground measurements, the method can be deployed at a large-scale and fine spatial resolution, and it enables comparative analyses among multiple cities.

Acknowledgements

We gratefully acknowledge the participants of the survey and the input data, and the the helpful comments received during the review process. We thank the members of the NUS Urban Analytics Lab for the discussions. The Institutional Review Board of the National University of Singapore has reviewed and approved the ethical aspects of this research (reference code NUS-IRB-2021-906). The first author has been supported by Department of Education of Guangdong Province. This research is part of the projects (i) National Science Foundation of China (42071360) and The Technology Project of The Science and Technology Commission of Shenzhen (JSGG20201103093401004); (ii) The Guangdong–Hong Kong-Macau Joint Laboratory Program, which is supported by The Guangdong Science and Technology Strategic Innovation Fund Grant 2020B1212030009; (iii) Large-scale 3D Geospatial Data for Urban Analytics, which is supported by the National University of Singapore under the Start Up Grant R-295-000-171-133.

and (iv) Shenzhen Key Laboratory of Digital Twin Technologies for Cities Grant ZDSYS20210623101800001.

References

Aiello, L.M., Schifanella, R., Quercia, D., Aletta, F., 2016. Chatty maps: constructing sound maps of urban areas from social media data. Royal Society open science 3, 150690.

Andringa, T.C., Lanser, J.J.L., 2013. How pleasant sounds promote and annoying sounds impede health: A cognitive approach. International journal of environmental research and public health 10, 1439–1461.

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google street view: Capturing the world at street level. Computer 43, 32–38.

Arietta, S.M., Efros, A.A., Ramamoorthi, R., Agrawala, M., 2014. City forensics: Using visual elements to predict non-visual city attributes. IEEE transactions on visualization and computer graphics 20, 2624–2633.

Axelsson, Ö., 2015. How to measure soundscape quality, in: Proceedings of the Euronoise 2015 conference, pp. 1477–1481.

Axelsson, Ö., Nilsson, M.E., Berglund, B., 2010. A principal components model of soundscape perception. The Journal of the Acoustical Society of America 128, 2836–2846.

Aytar, Y., Vondrick, C., Torralba, A., 2016. Soundnet: Learning sound representations from unlabeled video. Advances in neural information processing systems 29.

Becker, M., Caminiti, S., Fiorella, D., Francis, L., Gravino, P., Haklay, M., Hotho, A., Loreto, V., Mueller, J., Ricchiuti, F., et al., 2013. Awareness and learning in participatory noise sensing. PloS one 8, e81638.

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

Brooks, B.M., Schulte-Fortkamp, B., Voigt, K.S., Case, A.U., 2014. Exploring our sonic environment through soundscape research & theory. Acoustics Today 10, 30–40.

Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS Journal of Photogrammetry and Remote Sensing 163, 82–97.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Chen, W., Wu, A.N., Biljecki, F., 2021. Classification of urban morphology with deep learning: Application on urban vitality. Computers, Environment and Urban Systems 90, 101706.

Cheng, L., Chu, S., Zong, W., Li, S., Wu, J., Li, M., 2017. Use of tencent street view imagery for visual perception of streets. ISPRS International Journal of Geo-Information 6, 265.

Chew, Y.R., Wu, B.S., 2016. A soundscape approach to analyze traffic noise in the city of taipei, taiwan. Computers, Environment and Urban Systems 59, 78–85.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3213–3223.

Davies, W.J., Bruce, N.S., Murphy, J.E., 2014. Soundscape reproduction and synthesis. Acta Acustica United with Acustica 100, 285–292.

De Silva, C., Warusavitharana, E.J., Ratnayake, R., 2017. An examination of the temporal effects of environmental cues on pedestrians' feelings of safety. Computers, Environment and Urban Systems 64, 266–274.

Derksen, S., Keselman, H.J., 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology 45, 265–282.

Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A., 2016. Deep learning the city: Quantifying urban perception at a global scale, in: European conference on computer vision, Springer. pp. 196–212.

Einhäuser, W., da Silva, L.F., Bendixen, A., 2020. Intraindividual consistency between auditory and visual multistability. Perception 49, 119–138.

Fan, C., Esparza, M., Dargin, J., Wu, F., Oztekin, B., Mostafavi, A., 2020. Spatial biases in crowdsourced data: Social media content attention concentrates on populous areas in disasters. Computers, Environment and Urban Systems 83, 101514.

Friedman, J.H., 2002. Stochastic gradient boosting. Computational statistics & data analysis 38, 367–378.

Gasco, L., Clavel, C., Asensio, C., de Arcas, G., 2019. Beyond sound level monitoring: Exploitation of social media to gather citizens subjective response to noise. Science of The Total Environment 658, 69–79.

Gasco, L., Schifanella, R., Aiello, L.M., Quercia, D., Asensio, C., de Arcas, G., 2020. Social media and open data to quantify the effects of noise on health. Frontiers in Sustainable Cities 2, 41.

Guan, F., Fang, Z., Wang, L., Zhang, X., Zhong, H., Huang, H., 2022. Modelling people's perceived scene complexity of real-world environments using street-view panoramas and open geodata. ISPRS Journal of Photogrammetry and Remote Sensing 186, 315–331.

Hanibuchi, T., Nakaya, T., Inoue, S., 2019. Virtual audits of streetscapes by crowdworkers. Health & Place 59, 102203.

Harvey, C., Aultman-Hall, L., Hurley, S.E., Troy, A., 2015. Effects of skeletal streetscape design on perceived safety. Landscape and Urban Planning 142, 18–28.

Hasegawa, Y., Lau, S.K., 2022. Comprehensive audio-visual environmental effects on residential soundscapes and satisfaction: Partial least square structural equation modeling approach. Landscape and Urban Planning 220, 104351.

Hawes, J.K., Gounaridis, D., Newell, J.P., 2022. Does urban agriculture lead to gentrification? Landscape and Urban Planning 225, 104447.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Herzog, T.R., Kaplan, S., Kaplan, R., 1976. The prediction of preference for familiar urban places. Environment and Behavior 8, 627–645.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks

for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29, 82–97.

Hoffmann, B., Moebus, S., Stang, A., Beck, E.M., Dragano, N., Möhlenkamp, S., Schmermund, A., Memmesheimer, M., Mann, K., Erbel, R., et al., 2006. Residence close to high traffic and prevalence of coronary heart disease. European heart journal 27, 2696–2702.

Hong, J.Y., Jeon, J.Y., 2015. Influence of urban contexts on soundscape perceptions: A structural equation modeling approach. Landscape and Urban Planning 141, 78–87.

Hong, J.Y., Jeon, J.Y., 2017. Relationship between spatiotemporal variability of soundscape and urban morphology in a multifunctional urban area: A case study in seoul, korea. Building and Environment 126, 382–395.

Hsieh, H.P., Yen, T.C., Li, C.T., 2015. What makes new york so noisy? reasoning noise pollution by mining multimodal geo-social big data, in: Proceedings of the 23rd ACM international conference on Multimedia, pp. 181–184.

ISO/DIS 12913-1, 2014. Acoustics. soundscape–part 1: definition and conceptual framework.

Ito, K., Biljecki, F., 2021. Assessing bikeability with street view imagery and computer vision. Transportation Research Part C: Emerging Technologies 132, 103371. doi:10.1016/j.trc.2021.103371.

Jo, H.I., Jeon, J.Y., 2020. Effect of the appropriateness of sound environment on urban soundscape assessment. Building and environment 179, 106975.

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. doi:10.1016/j.cities.2021.103333.

Korpilo, S., Nyberg, E., Vierikko, K., Nieminen, H., Arciniegas, G., Raymond, C.M., 2023. Developing a multi-sensory public participation gis (msppgis) method for integrating landscape values and soundscapes of urban green infrastructure. Landscape and Urban Planning 230, 104617.

Kruse, J., Kang, Y., Liu, Y.N., Zhang, F., Gao, S., 2021. Places for play: Understanding human perception of playability in cities using street view images and deep learning. Computers, Environment and Urban Systems 90, 101693. doi:10.1016/j.compenvurbsys.2021.101693.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

Liu, J., Kang, J., Behm, H., Luo, T., 2014. Effects of landscape on soundscape perception: Soundwalks in city parks. Landscape and urban planning 123, 30–40.

Liu, P., Biljecki, F., 2022. A review of spatially-explicit geoai applications in urban geography. International Journal of Applied Earth Observation and Geoinformation 112, 102936.

Long, Y., Liu, L., 2017. How green are the streets? an analysis for central areas of chinese cities using tencent street view. PloS one 12, e0171110.

Mahabir, R., Schuchard, R., Crooks, A., Croitoru, A., Stefanidis, A., 2020. Crowdsourcing street view imagery: a comparison of mapillary and openstreetcam. ISPRS International Journal of Geo-Information 9, 341.

Min, W., Mei, S., Liu, L., Wang, Y., Jiang, S., 2019. Multi-task deep relative attribute learning for visual urban perception. IEEE Transactions on Image Processing 29, 657–669.

Nagata, S., Nakaya, T., Hanibuchi, T., Amagasa, S., Kikuchi, H., Inoue, S., 2020. Objective scoring of streetscape walkability related to leisure walking: Statistical modeling approach with semantic segmentation of google street view images. Health & Place 66, 102428.

Naik, N., Philipoom, J., Raskar, R., Hidalgo, C., 2014. Streetscore-predicting the perceived safety of one million streetscapes, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 779–785.

Nguyen, Q.C., Huang, Y., Kumar, A., Duan, H., Keralis, J.M., Dwivedi, P., Meng, H.W., Brunisholz, K.D., Jay, J., Javanmardi, M., et al., 2020. Using 164 million google street view images to derive built environment predictors of covid-19 cases. International journal of environmental research and public health 17, 6359.

Nilsson, M.E., Berglund, B., 2006. Soundscape quality in suburban green areas and city parks. Acta Acustica united with Acustica 92, 903–911.

Ning, H., Li, Z., Wang, C., Hodgson, M.E., Huang, X., Li, X., 2022. Converting street view images to land cover maps for metric mapping: a case study on

sidewalk network extraction for the wheelchair users. Computers, Environment and Urban Systems 95, 101808.

Ordonez, V., Berg, T.L., 2014. Learning high-level judgments of urban perception, in: European conference on computer vision, Springer. pp. 494–510.

Radicchi, A., Henckel, D., Memmel, M., 2016. Citizens as smart, active sensors for a quiet and just city. the case of the "open source soundscapes" approach to identify, assess and plan "everyday quiet areas" in cities. Noise mapping 5, 1–20.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28, 91–99.

Salem, T., Zhai, M., Workman, S., Jacobs, N., 2018. A multimodal approach to mapping soundscapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2524–2527.

Schafer, R.M., 1993. The soundscape: Our sonic environment and the tuning of the world. Simon and Schuster.

Shi, W., 2021. Introduction to urban sensing, in: Urban Informatics. Springer, pp. 311–314.

Song, G., Liu, L., He, S., Cai, L., Xu, C., 2020. Safety perceptions among african migrants in guangzhou and foshan, china. Cities 99, 102624.

Stansfeld, S.A., Berglund, B., Clark, C., Lopez-Barrio, I., Fischer, P., Öhrström, E., Haines, M.M., Head, J., Hygge, S., Van Kamp, I., et al., 2005. Aircraft and road traffic noise and children's cognition and health: a cross-national study. The Lancet 365, 1942–1949.

Stoter, J., Peters, R., Commandeur, T., Dukai, B., Kumar, K., Ledoux, H., 2020. Automated reconstruction of 3D input data for noise simulation. Computers, Environment and Urban Systems 80, 101424.

Sudarsono, A.S., Lam, Y.W., Davies, W.J., 2017. The validation of acoustic environment simulator to determine the relationship between sound objects and soundscape. Acta Acustica united with Acustica 103, 657–667.

Tang, J., Long, Y., 2019. Measuring visual quality of street space and its temporal variation: Methodology and its application in the hutong area in beijing. Landscape and Urban Planning 191, 103436.

Tong, H., Kang, J., 2021. Relationships between noise complaints and socio-economic factors in england. Sustainable Cities and Society 65, 102573.

Tu, W., Zhu, T., Xia, J., Zhou, Y., Lai, Y., Jiang, J., Li, Q., 2020. Portraying the spatial dynamics of urban vibrancy using multisource urban big data. Computers, Environment and Urban Systems 80, 101428.

Van Renterghem, T., 2019. Towards explaining the positive effect of vegetation on the perception of environmental noise. Urban Forestry & Urban Greening 40, 133–144.

Van Renterghem, T., Vanhecke, K., Filipan, K., Sun, K., De Pessemier, T., De Coensel, B., Joseph, W., Botteldooren, D., 2020. Interactive soundscape augmentation by natural sounds in a noise polluted urban park. Landscape and urban planning 194, 103705.

Verma, D., Jana, A., Ramamritham, K., 2019. Artificial intelligence and human senses for the evaluation of urban surroundings, in: International Conference on Intelligent Human Systems Integration, Springer. pp. 852–857.

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, M., Vermeulen, F., 2021. Life between buildings from a street view image: What do big data analytics reveal about neighbourhood organisational vitality? Urban Studies 58, 3118–3139.

Wang, R., Liu, Y., Lu, Y., Zhang, J., Liu, P., Yao, Y., Grekousis, G., 2019. Perceptions of built environment and health outcomes for older chinese in beijing: A big data approach with street view images and deep learning technique. Computers, Environment and Urban Systems 78, 101386.

Wu, A.N., Biljecki, F., 2021. Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability. Landscape and Urban Planning 214, 104167.

Wu, A.N., Biljecki, F., 2022. GANmapper: geographical data translation. International Journal of Geographical Information Science 36, 1394–1422.

Wu, D., Gong, J., Liang, J., Sun, J., Zhang, G., 2020. Analyzing the influence of urban street greening and street buildings on summertime air pollution based on street view image data. ISPRS International Journal of Geo-Information 9, 500.

Yan, Y., Feng, C.C., Huang, W., Fan, H., Wang, Y.C., Zipf, A., 2020. Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience. International Journal of Geographical Information Science 34, 1–27.

Yao, Y., Liang, Z., Yuan, Z., Liu, P., Bie, Y., Zhang, J., Wang, R., Wang, J., Guan, Q., 2019. A human-machine adversarial scoring framework for urban perception assessment using street-view images. International Journal of Geographical Information Science 33, 2363–2384.

Yong Jeon, J., Jik Lee, P., Young Hong, J., Cabrera, D., 2011. Non-auditory factors affecting urban soundscape evaluation. The Journal of the Acoustical Society of America 130, 3761–3770.

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

Zhang, F., 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, J., Fukuda, T., Yabuki, N., 2021b. Development of a city-scale approach for façade color measurement with building functional classification using deep learning and street view images. ISPRS International Journal of Geo-Information 10, 551.

Zhang, Y., Li, Q., Tu, W., Mai, K., Yao, Y., Chen, Y., 2019. Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. Computers, Environment and Urban Systems 78, 101374.

Zhao, Y., Sheppard, S., Sun, Z., Hao, Z., Jin, J., Bai, Z., Bian, Q., Wang, C., 2022. Soundscapes of urban parks: An innovative approach for ecosystem monitoring and adaptive management. Urban Forestry & Urban Greening 71, 127555.

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, 1452–1464.