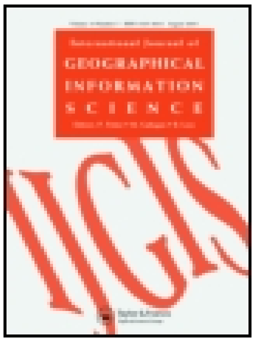

Taylor & Francis
Taylor & Francis Group

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

A human-machine adversarial scoring framework for urban perception assessment using street-view images

Yao Yao, Zhaotang Liang, Zehao Yuan, Penghua Liu, Yongpan Bie, Jinbao Zhang, Ruoyu Wang, Jiale Wang & Qingfeng Guan

To cite this article: Yao Yao, Zhaotang Liang, Zehao Yuan, Penghua Liu, Yongpan Bie, Jinbao Zhang, Ruoyu Wang, Jiale Wang & Qingfeng Guan (2019): A human-machine adversarial scoring framework for urban perception assessment using street-view images, International Journal of Geographical Information Science

To link to this article: https://doi.org/10.1080/13658816.2019.1643024

Published online: 19 Jul 2019.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2019.1643024

RESEARCH ARTICLE

A human-machine adversarial scoring framework for urban perception assessment using street-view images

Yao Yao[a,b], Zhaotang Liang[c], Zehao Yuan[a], Penghua Liu[d], Yongpan Bie[a], Jinbao Zhang[d,e], Ruoyu Wang[d,f], Jiale Wang[g] and Qingfeng Guan[a]

[a]School of Geography and Information Engineering, China University of Geosciences, Wuhan, Hubei, China; [b]Alibaba Group, Hangzhou, Zhejiang, China; [c]Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong; [d]School of Geography and Planning, Sun Yat-sen University, Guangzhou, Guangdong, China; [e]Tencent Technology Inc., Shenzhen, Guangdong, China; [f]Institute of Geography, School of GeoSciences, University of Edinburgh, Edinburgh, UK; [g]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, Hubei, China

ABSTRACT

Though global-coverage urban perception datasets have been recently created using machine learning, their efficacy in accurately assessing local urban perceptions for other countries and regions remains a problem. Here we describe a human-machine adversarial scoring framework using a methodology that incorporates deep learning and iterative feedback with recommendation scores, which allows for the rapid and cost-effective assessment of the local urban perceptions for Chinese cities. Using the state-of-the-art Fully Convolutional Network (FCN) and Random Forest (RF) algorithms, the proposed method provides perception estimations with errors less than 10%. The driving factor analysis from both the visual and urban functional aspects demonstrated its feasibility in facilitating local urban perception derivations. With high-throughput and high-accuracy scorings, the proposed human-machine adversarial framework offers an affordable and rapid solution for urban planners and researchers to conduct local urban perception assessments.

ARTICLE HISTORY
Received 28 August 2018
Accepted 9 July 2019

KEYWORDS
Street view; urban perception; deep learning; urban planning; human-machine adversarial scoring

1. Introduction

Urban perceptions, which are the psychological feelings held by residents about an urban locale (Tuan 2013, Ordonez and Berg 2014), provide an important basis for understanding the ways in which urban environments interact with public mental health (Ulrich 1979, Frank and Engelke 2001, Wolch *et al*. 2014). Traditionally, the evaluation of human perceptions towards their visual surroundings remains difficult due to the lack of high-throughput methods, inadequate sample problems and being restricted to interviews and questionnaires (Hannay 1983, Halpern 1995, Kabisch *et al*. 2015, Dadvand *et al*. 2016). Given the costly and time-consuming nature of these investigation methods, a framework that can boost work efficiency is needed to optimize the urban perception assessment process.

CONTACT Qingfeng Guan guanqf@cug.edu.cn

The last several years have witnessed the fast development in multi-sources of geospatial big data, especially the emergence of massive geo-tagged imagery datasets (e.g., street-view imagery and check-in imagery) (Zhou *et al*. 2014). Such images with geographic location information contain abundant visual information (Xu *et al*. 2017) and thus could effectively reflect visual scenery that is seen in daily life (Hu *et al*. 2015). Since a sight is the most intuitive way for urban residents to gain perceptions about their surrounding environments (Ulrich 1979), geo-tagged image datasets offer new opportunities to tackle the large-scale derivation problem for urban perception. Street-view (SV) imagery, which is composed of panoramic views from various positions along streets, has emerged as a promising data source to infer urban perceptions. Street-view pictures are geo-tagged photos that are collected, processed and maintained by map service providers (e.g., Google Maps and Tencent Maps) using a standard processing method, and they are collected through dedicated devices by acquiring images from different headings and pitches (Anguelov *et al*. 2010). Street-view imagery is primarily distributed along urban streets (Cheng *et al*. 2017) and represents the physical morphological properties of urban interior spaces (Gebru *et al*. 2017, Zhang *et al*. 2018a).

Salesses *et al*. (2013) first proposed using street-view images to assess the effect of a city's environment on social and economic outcomes by collecting human perceptions through pair-wise street-view image comparisons. Based on Salesses's work, Dubey *et al*. (2016) extended the surveying area to global major cities using an online crowdsourcing strategy and machine learning in computer vision to build a large-scale urban perception global dataset, thus overcoming the inadequate sample problem and certain limits imposed by traditional interview and questionnaire approach. To the best of our knowledge, the existing studies assessed urban perceptions based on the dataset provided by the MIT Place Pulse project (Place Pulse 1.0 and 2.0) (Ordonez and Berg 2014, Porzi *et al*. 2015, Naik *et al*. 2017, Zhang *et al*. 2018a, 2018b).

Though global-coverage urban perception datasets have been created by Dubey *et al*. (2016) using machine learning, its efficacy in accurately assessing local urban perceptions remains a problem. For example, training sample areas in this dataset only contain two Chinese regions (Hong Kong and Taiwan). Previous studies indicate that Hong Kong and Taiwan are largely different from mainland China in terms of their special political and economic status and physical environments (Wong 2015). The urban perceptions derived from such a global dataset may not be representative of cities in mainland China. Due to the complexity of China's urban and local environments, applications of models that are trained by Place Pulse dataset (which mainly consists of western scenes) to a Chinese city are problematic.

Chinese cities manifest different environments from other cities in the world, similar to how Eastern European cities (post-communist ones) differ from the American cities or Australian cities. Previous studies showed differences of Eastern and Western architectures and town planning style in the 'demarcations' of interior and exterior as well as private and public spaces through discussions of differences in street-view images (Ashihara 1983). Every city is a complex system, composed of people, places, routes and activities distinctive from cities of other countries (Cameron and Larsen-Freeman 2007). People in a Chinese city commonly feel safe about their inner city regardless of the city's physical appearance, since the inner city is usually densely populated and has more police officers.

Inner-city images in China may look like urban villages – not good-looking and disordered – but they would be perceived as being quite safe by Chinese people.

Also, an urban perception is a subjective assessment and is influenced by people's social and cultural backgrounds (Rapoport and Hawkes 1970). Although Dubey *et al*. (2016) and Salesses *et al*. (2013) claimed that demographics of the survey respondents would not cause any bias, their statement is only tenable within sample areas or cities that are similar to the sample areas, and bias may still exist when this dataset is directly used to predict and assess urban perceptions elsewhere. In other words, when we need to accurately assess the urban perception of a region, we need to obtain a local urban perception dataset from the residents who are aware of the regional socioeconomic background.

To address these problems on assessing local urban perceptions, we proposed a novel 'human-machine adversarial' scoring methodology to rapidly and cost-effectively assess local urban perceptions. This study developed a framework with deep learning, street-view imagery and iterative feedback mechanism and to assess city-scale urban perceptions. We conducted a case study of an urban perception assessment in a high-density urban environment, e.g., Wuhan, to demonstrate the efficacy of the proposed framework. Moreover, we analyzed the driving factors to explain the results from both the visual and urban functional aspects.

2. Methodology

The flowchart of the proposed methodology is illustrated in Figure 1. The methodology attempts to assess the local urban perception using the proposed human-machine adversarial framework. The framework includes three procedural components: 1) using fully connection network (FCN) trained by the ADE-20K[1] dataset to semantically segment features in each street-view photograph and obtain the areal ratio of each semantic object; 2) using the proposed human-machine adversarial scoring module to enhance the efficiency in the urban perception assessment for Chinese cities; and 3) analyzing the driving factors of the derived urban perceptions in term of the visual elements and urban functions.

2.1. *Designing 'Human-machine adversarial' strategy*

Humans have superior abilities to recognize image's global-property, which provides the theoretical support for our 'human-machine adversarial' methodology (Greene and Oliva 2009a). Previous experimental comparisons indicate that global-property categorization takes significantly less presentation time than basic-level categorization, for example, the degree of openness or navigable, rather than a mountain or lake. Our visual system can recognize and classify scenes much faster than individual component objects, which usually take less about 100 ms in laboratory conditions (Greene and Oliva 2009b).

Human perceptions (e.g., safety, lively, etc.) on street-view scenes – scenes that we see daily – are exactly a dual to global-property for natural scenes (e.g., openness, temperature, etc.). Therefore, human's superiority for understanding global-property is used to facilitate human annotation process for human perception (global-property) ratings in this study. The 'human-machine adversarial' design provides a conditioned reflex environment where machine learning assists human annotations on global-property categorization, similar to Google's recent research 'Fluid Annotation (2018)'.

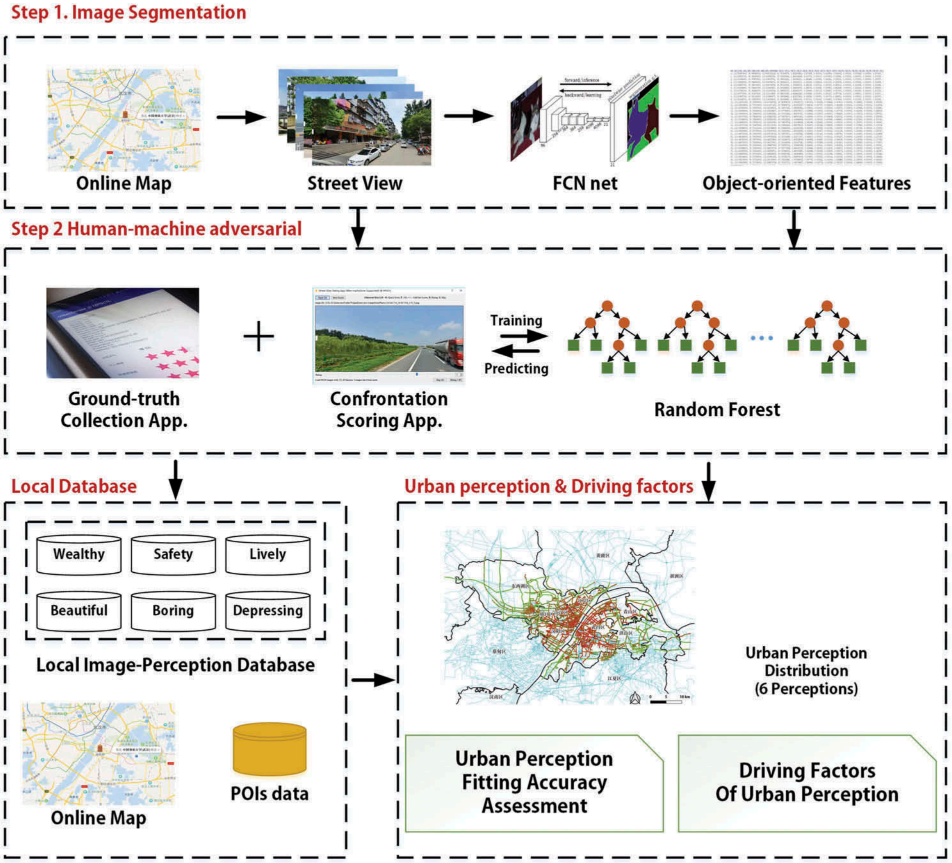
Step 1. Image Segmentation
Online Map
Street View
FCN net
Object-oriented Features
Step 2 Human-machine adversarial
Ground-truth Collection App.
Confrontation Scoring App.
Training
Predicting
Random Forest
Local Database
Wealthy
Safety
Lively
Beautiful
Boring
Depressing
Local Image-Perception Database
Online Map
POIs data
Urban perception & Driving factors
Urban Perception Distribution (6 Perceptions)
Urban Perception Fitting Accuracy Assessment
Driving Factors Of Urban Perception

Figure 1. Workflow of assessing the street-level local urban perception via human-machine adversarial scoring.

2.2. Human and machine rating local street view images

2.2.1. Semantic segmentation of street-view imagery via FCN-8s

To determine the visual elements that might induce a safe, lively, or depressing perception of a place, our proposed method aims to extract the semantic elements of a place that might be highly correlated with human perceptions. The extracted semantic elements then proceed for the RF-model fitting in adversarial scoring and further correlation analysis. Recent progresses on deep learning show a fully convolutional network (FCN) can predict each pixel's semantic property in an image, which can be used to produce natural-object-level segmentation results.(Long *et al*. 2015, Badrinarayanan *et al*. 2017).

As shown in Figure 2(a), a FCN divides a street-view image into multiple sub-scenes, each of which attends to vehicles, roads, trees or other natural objects up to 151 categories (including the category of 'unknown'). The MIT ADE20K website[2] documents the full description of all 151 categories. In this study, we use the ADE20K dataset released by MIT (Zhou *et al*. 2019, 2017) to train our FCN network. Next, the trained FCN network is integrated into our proposed human-machine adversarial scoring framework,

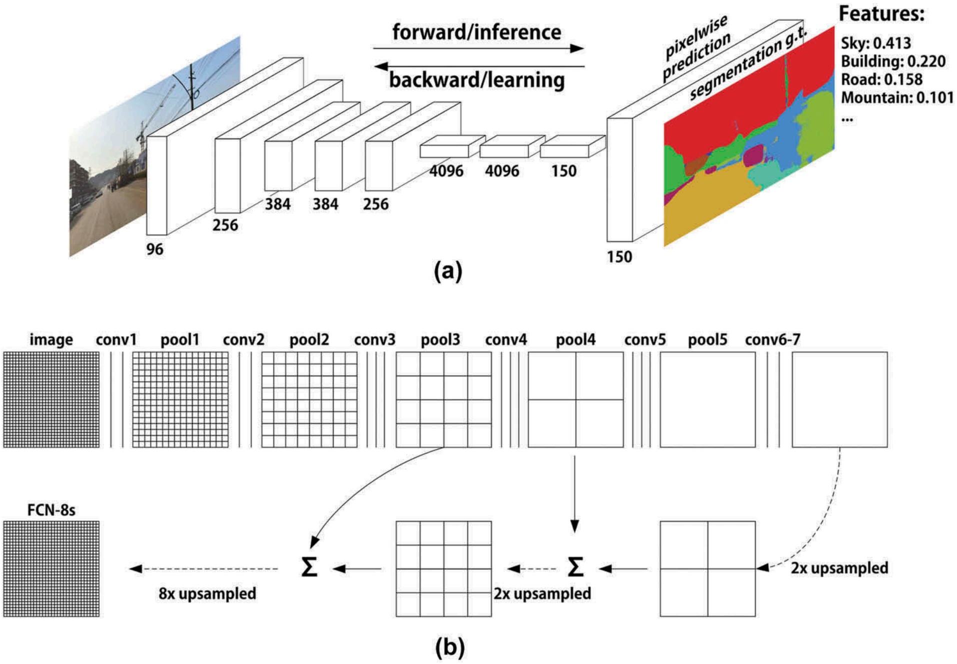

Figure 2. (a) The input and output of the fully convolutional network (FCN) and (b) the details of the FCN structure (Long *et al*. 2015).

which provides a 151-dimension feature vector for the random forest to fit the human-annotators' scoring preferences and give recommendation scores.

Coupled with the calculation of the per-pixel loss based on the softmax layer (Shelhamer *et al*. 2014, Zheng *et al*. 2016), the FCN-8s network (Figure 2(b)) produces the area ratio of each visual element in the image by counting the number of pixels in each segmentation mask.

2.2.2. Urban perception derivation via the human-machine adversarial scoring framework

This study focuses on six categories of urban perceptions: wealthy, safety, lively, beautiful, boring and depressing, as in Place Pulse 2.0 (Dubey *et al*. 2016). These six perception datasets from previous studies are assembled through training on mostly western urban street scenery with CNNs and annotators' votes on pairwise image comparisons (Dubey *et al*. 2016). Our proposed framework, however, collects real score annotations on each image from volunteers. To accurately produce urban perceptions in Chinese urban environments, local volunteers who are aware of the regional socioeconomic background are asked to work through the human-machine adversarial scoring framework.

Our human-machine adversarial scoring module processed the FCN sematic segmentation result and the city-scale human perceptions from local volunteers. Furthermore, the module mined the relationship between the visual scenery and perception directly to expedite image classification according to human perception. We use RF algorithm to determine the final image classification, and therefore the results are subject to RF fitting and limited to one perception label per image.

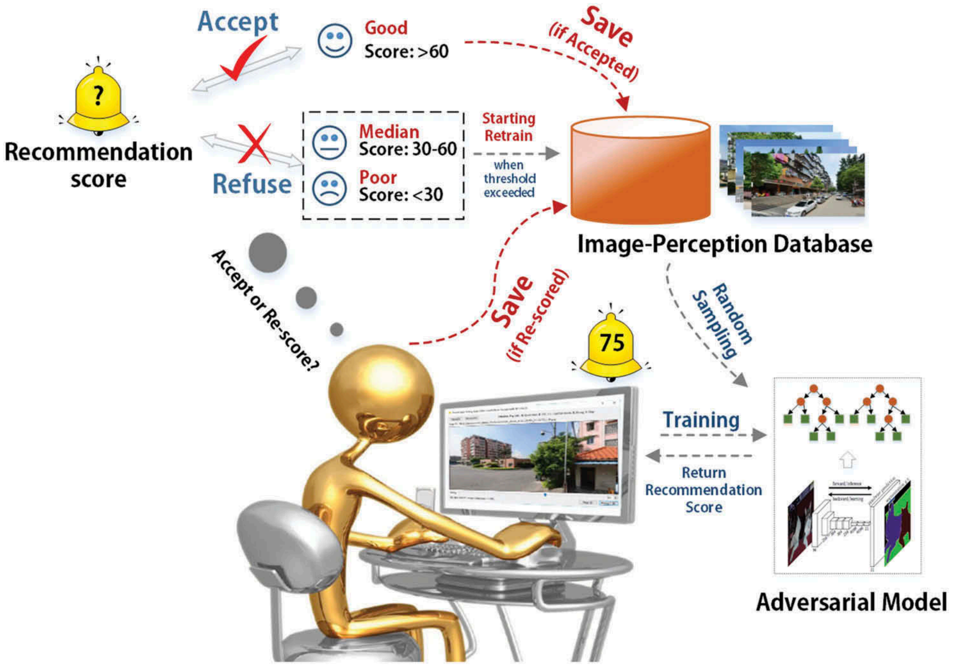

Figure 3. Schematic diagram of the proposed human-machine adversarial scoring process.

Figure 3 illustrates the process of human-machine adversarial scoring. Volunteers score a displayed street-view image in terms of the six types of perception in a range of 1–100 with 0 being the lowest and 100 being the highest level of a perception. The demography and number of volunteers recruited to score street-view images may vary depending on the different application cases. A case study is provided in Section 4 to demonstrate the feasibility and effectiveness of the proposed method.

2.2.2.1. Random-forest fitting. The proposed method consists of a random forest (RF)-based module to fit the relationship between the visual scenic features and the user scorings. The visual scenic feature is a 151-dimension vector that reflects the areal proportion of each kind of object in the FCN segmentation. Once a user has scored the first 50 photos, the scoring software establishes a random forest set to fit the scoring process. Then, as users have rated subsequent photos, the software offers a recommendation score based on the rules learned from the previous user rating actions. Previous studies have already demonstrated RFs' outstanding performance in model fitting (Fern A Ndez-Delgado *et al*. 2014).

During the random forest training process, the training data set is randomly divided into a training (in-bag) data set and a test (out-of-bag, OOB) data set. The OOB data set is only used to test the model accuracy at each iteration during the training process. The average OOB validation error can be used to evaluate the degree to which the RF-based fit or classification model achieves the best accuracy. Previous studies have proven that the OOB estimation is better than the cross-validation (Fern A Ndez-Delgado *et al*. 2014).

2.2.2.2. Iterative adjusting. Inspired by the iterative feedback module designed for scoring cell phenotypes (Jones *et al.* 2009), we enable our scoring software to automatically adjust the recommendation scores according to the user scoring behaviors. If the recommendation scores of more than five pictures seriously deviate from a user's score by more than 10 points, the embedded random forest module will be retrained and self-correct the fitting model. Otherwise, if the OOB validation error of the fitting model is less than 10 points, the user scoring procedure stops and outputs a human-machine adversarial scoring dataset.

The adversarial scoring between people and machines help obtain a stable machine learning model while human-machine confrontation reaches a compromise. We define Human-machine confrontation compromise as a scoring result that difference of machine prediction score and human annotation score is within ±5 point. As the photos assigned to volunteers are randomly retrieved from the after-segmentation street-view image database, if a photograph is rated multiple times by several volunteers, the final score of the photograph will be set as the median value to avoid extreme scores. The final product is a data set of scored local perceptions on street-view imagery ready for analyzing urban perceptions of Chinese cities.

2.2.3. Accuracy assessment

During the process of human-machine adversarial scoring fitting via random forest and the process of RF-fitting between urban perception and POI-based urban function, this study uses the Pearson correlation coefficient (Pearson R), standard R^2, root mean squared error (RMSE) and mean absolute error (MAE) to quantify the accuracy between the predictions and the ground-truth values. The Pearson R, standard R^2, RMSE and MAE are mathematically represented as follows using Equation (1) to Equation (4), respectively.

$$\textit{Pearson } R = \frac{\sum_{i=1}^{n}(y_i - \overline{y_i})\left(\widehat{y_i} - \overline{\widehat{y_i}}\right)}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}\sqrt{\sum_{i=1}^{n}\left(\widehat{y_i} - \overline{\widehat{y_i}}\right)^2}} \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

Where y_i is the ground-truth value, $\bar{y}$ is equal to $\frac{1}{n}\sum_{i=1}^{n} y_i$, and $\hat{y}_i$ is the predicted result from the fitting model.

2.3. Exploring the driving factors of urban perceptions

The data set of local perception scores can be used to estimate the perceptual distributions at the city-scale. We use the trained RF model embedded in the adversarial scoring module to estimate all four headings of the street-view pictures (0, 90, 180 and 270 degrees) and calculate the average perception score at each site. Next, we assemble the average scores for all sites to map the urban perception for a given study area.

Driving factor analyses aim to further assess the derived urban perception from two aspects. The first aspect is semantic element factor identification. The weights of the features extracted from the RF training process provide the basis to quantify the effects of street level sub-scenes on urban perceptions as a means to examine the rationality of the derived urban perception (Zhang *et al*. 2018c). The second aspect is POI-based urban function factor identification. A detailed RF-based calculation method from previous studies (Palczewska *et al*. 2014, Yao *et al*. 2017a, 2017b) is used to analyze the relationship between urban functional patterns and urban perceptions. The POI-based identification supplements the semantic element analysis result in assessing the feasibility and effectiveness of the proposed human-machine adversarial framework for urban perception assessment.

The spatial distribution of POIs is used to construct an RF-based model that fits the urban perception distribution. We use the fitting accuracy and parameters to calculate the correlation between the urban functional patterns and urban perceptions.

3. Case study

Taking Wuhan as the case study, we performed the human-machine adversarial scoring procedure and conducted driving factor analysis to show that the proposed methodology is an efficient and low-cost method for obtaining city-scale local urban perception.

As the largest city in and the political, economic and cultural center of Central China (Sun *et al*. 2016, Yao, Wang *et al*. 2016), Wuhan is characterized as one of the most rapidly developing cities in China. This case study covers the downtown area of Wuhan, including 8 administrative districts[3] (Figure 4). The central zone of Wuhan, including Wuchang, Jiang'an, Jianghan and Hanyang, along the Yangtze River is the most developed area of Wuhan.

Street-view (SV) images are essential data in this study. Tencent Maps (https://map.qq.com/) is one of the largest online map service providers in China (Long and Liu 2017). Similar to Google Maps, Tencent Maps provides street-view photos (Figure 5) for various positions with different headings and pitches along each road. Based on the road network data (Figure 4) from OpenStreetMap.org, we evenly selected our sampling points 100 meters apart on every main road. In our sampling strategy, each sample point captures street-view images from four headings (0, 90, 180, and 270 degrees) with a fixed horizontal pitch, as illustrated in Figure 5. In total, we collected nearly 500 thousands street-view photographs of major Chinese cities (Beijing, Shenzhen, Guangzhou, Shanghai, Wuhan, Hangzhou, etc.). In terms of our case study in Wuhan, we selected 24,860 sampling points and obtained a total of 99,440 street-view images for further processing.

In addition, we fetched Gaode POI data to analyze the relationship between urban functions and urban perceptions in this case study. Gaode is one of the biggest map service providers in China and has a rich source of POI (http://amap.com). We captured

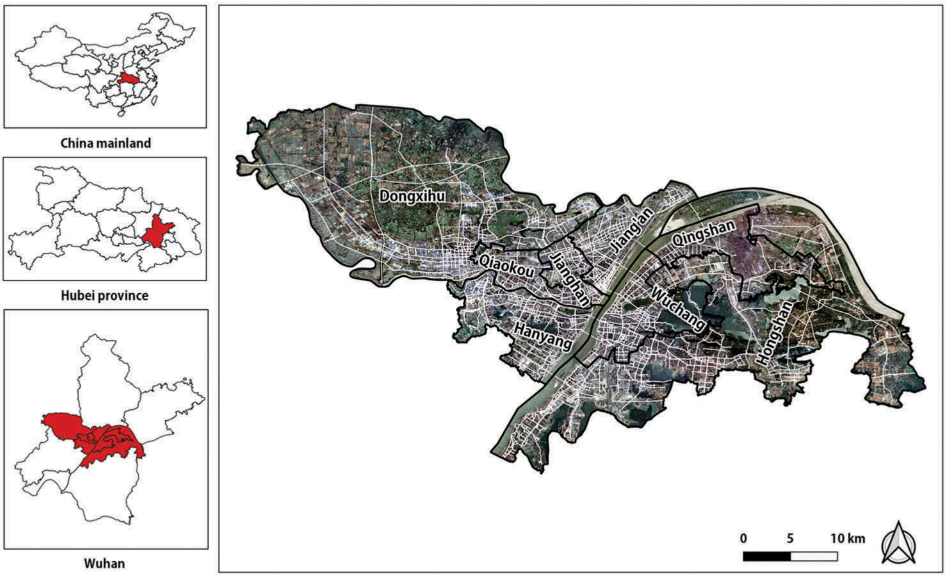

Figure 4. Case study area: Downtown area of Wuhan, Hubei Province. The white lines in the right subplot are the main roads in the study area obtained from openstreetmap.org.

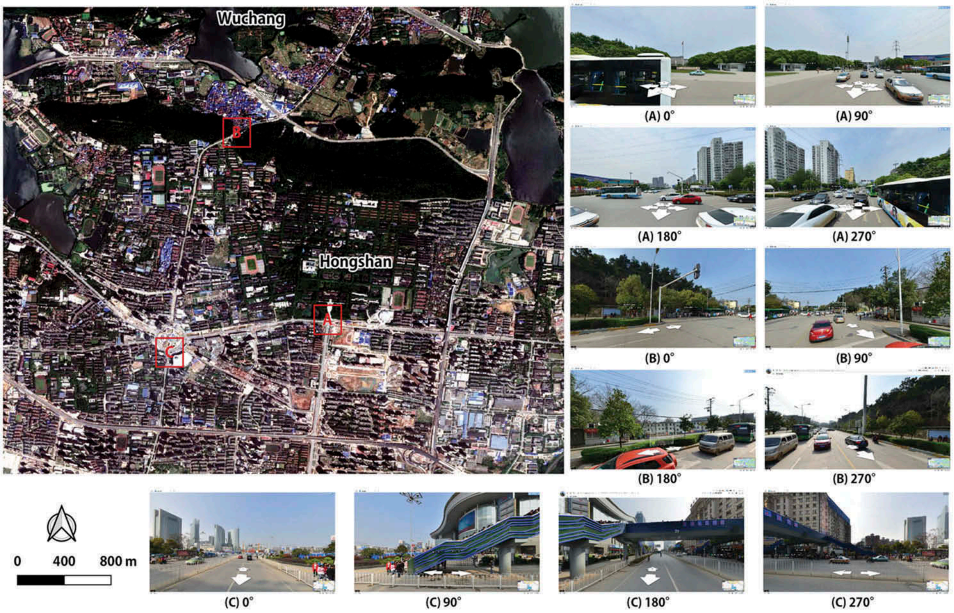

Figure 5. Online Tencent street-view data. Case areas: (a) Huazhong university of science and technology, (b) Nanwang mountain, and (c) Wuhan optical valley (CBD area).

603,015 POIs from Gaode Maps (https://www.amap.com/) in the case study area. Gaode provides nine categories of POI, and this data source was successfully employed in the recognition of urban function patterns (Liu *et al*. 2017, Yao *et al*. 2018). Figure 6 shows the density distribution of the nine POI categories.

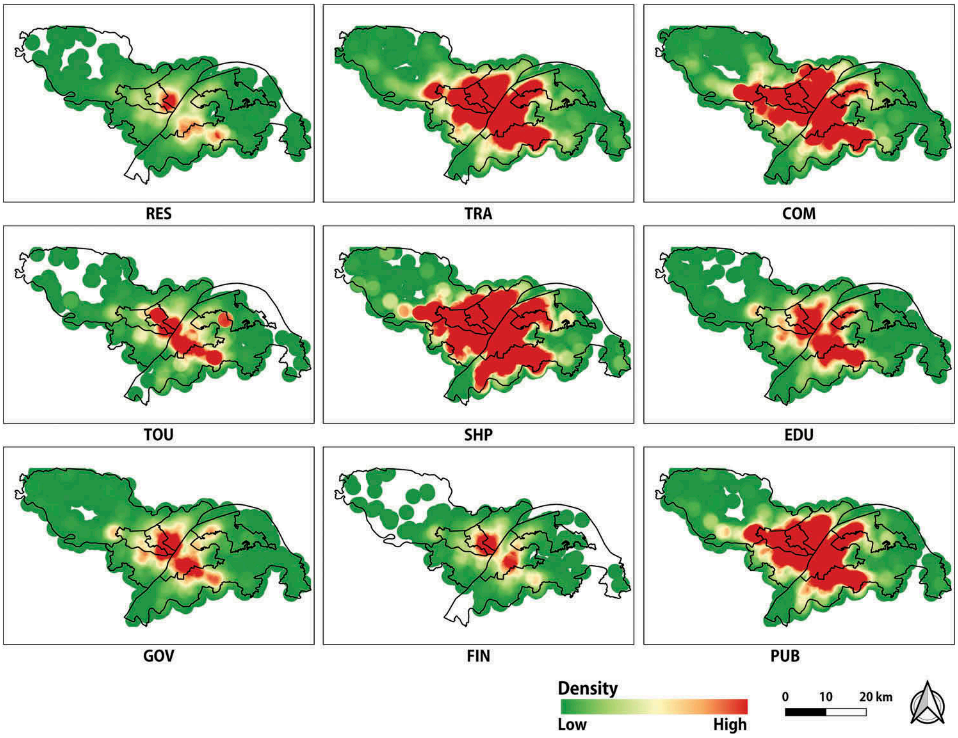

Figure 6. Distribution of Gaode POIs in the study area using kernel density. POI categories: residential communities (RES), traffic facilities (TRA), commercial and business (COM), tourist attractions (TOU), food and shopping (SHP), education facilities (EDU), government and public services (GOV), financial services (FIN), and public facilities (PUB).

4. Results

Our research team built a software application and developed the proposed models in Section 3. The FCN model was trained using two sets of Nvidia GTX 1080ti graphics processors. Several open-source C/C++ libraries, such as GDAL (http://www.gdal.org/), DLib (http://dlib.net/), Qt (https://www.qt.io/) and Shark (http://image.diku.dk/shark/), were used in this project. The source codes were implemented in Java, C++ with OpenMP and run on a multiprocessor computational server. The related application and source code can be downloaded from our GitHub website (https://github.com/whuyao/human-machine-adversarial).

4.1. Semantic segmentation result

The ADE-20K dataset had a total of 20,210 items of training data and 2,000 items of validation data. When training the FCN-8s, the scanning window was set to 500*500 pixels, while the learning rate and the early-stopping minimum learning rate were set to 0.1 and 0.001, respectively. The batch size for each input was set to 32. To derive an FCN-8s suitable for semantic segmentation, nearly 1 week was required to complete the training procedure.

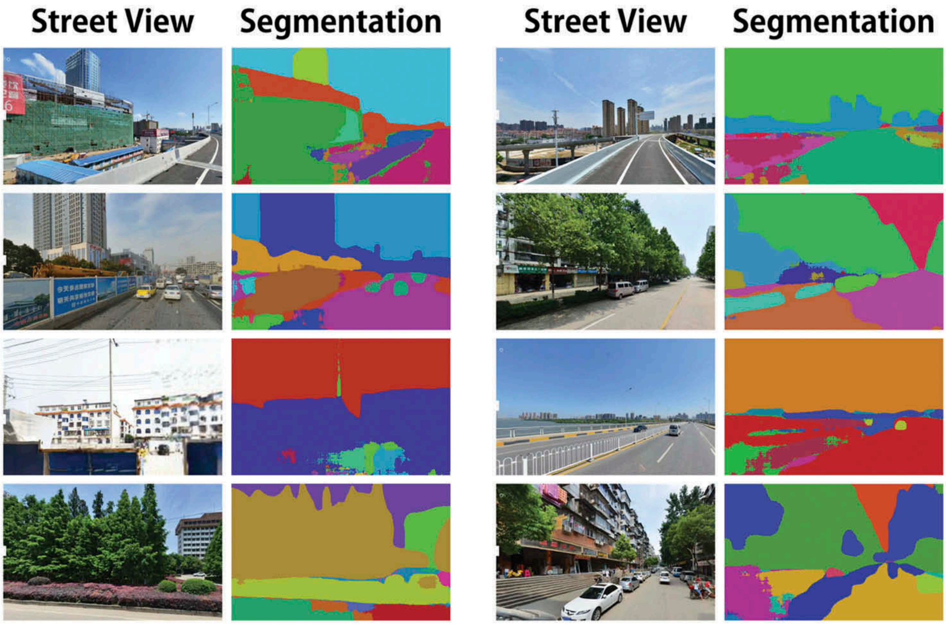

Figure 7. Scene segmentation results of the Tencent street-view photos via our trained FCN-8s. The colors of segmentation masks are random.

The trained FCN-8s showed excellent performance for natural image scene segmentation. Through a pixel-by-pixel comparison, experiments using the ADE-20K dataset showed that our trained FCN-8s achieved an accuracy of 81.44% on the training dataset and 66.83% on the test dataset. Figure 7 shows the segmentation results of the street-view images in the case study area. The street-level sceneries of Chinese cities are quite complex, ranging from under-construction environments to both rural and developed urban areas. However, the model trained in this paper showed satisfactory results in segmenting and parsing street-level sub-scenes, thereby handling the complexity of China's city environment quite well.

4.2. Human-machine adversarial scoring result

We invited a total of 20 college students and staff to be volunteers, and their ages ranged from 20 to 50 years old. The age distribution of volunteers is as follows: 9 people from 20 to 30 years old, 8 people from 31 to 40 years old, and 3 people from 41 to 50 years old. The ratio of male to female was approximately 1:1. A total of 25,000 street-view images were scored via the proposed human-machine adversarial scoring.

With the support of the human-machine adversarial scoring system, each person annotated 1,000–2,000 images in one or two hours since the recommendation scores provided by the program accelerated the scoring process. Through the adversarial scoring process, the 'human-machine compromise' state reached after 1,000 images manually scored by volunteers upon the training of the best fitting random forest model was complete and the module terminated the data collection process.

4.2.1. Human-machine RF training accuracy

A total of 60% of the user-scored perceptual annotation data was used to train the random forests, while the other 40% was used as the testing data for model validation. The fitting results are shown in Table 1. The RF-based urban perception estimation accuracy was over 90% on average, thereby demonstrating the strength of the sub-scene descriptors using the semantic segmentation technique. In the case of a full mark of 100 points, our average fitting error for the machine scoring was within 10 points. Since we established the stop criterion of the user scoring procedure as OOB errors less than 10 points, the training accuracy result for the lively perceptions (±10.11) serves the baseline result. Compared with the wealthy, safety and depressing perceptions, the human perceptions of beautiful and boring had obviously poorer accuracies with scoring errors of (±14.52) and (±11.01), respectively. These results indicated a higher perceptual diversity of beautiful and boring among our volunteers, which may represent a stronger subjectivity towards beautiful and boring from the demographic groups with higher-education backgrounds in the Wuhan region.

4.2.2. Correlation analyses among perceptions

The correlation matrix between the pairwise perceptions is shown in Table 2. Wealthy perceptions have strong positive correlations with safety and lively perceptions. Since human perception of wealth is usually stronger in Chinese downtown areas where denser populations and more modern facilities (as well as more police forces) are located, it is reasonable that safety and lively perceptions would be simultaneously evoked. Along with the increased/decreased degrees of prosperity, positive correlations appear 'interlocking' among wealthy, safety and lively perceptions. This

Table 1. Training accuracy of the urban perception estimation via random forest.

Perceptions	Average error	RMSE	OOB Error	OOB RMSE
Wealthy	1.84%	3.00	5.38%	8.60
Safety	1.37%	2.58	3.97%	7.32
Lively	2.36%	3.48	6.92%	10.11
Beautiful	3.88%	5.06	11.37%	14.52
Boring	2.61%	3.78	7.77%	11.01
Depressing	2.02%	2.91	5.99%	8.57

Table 2. The Pearson correlation matrix among the different perceptions.

Perceptions	Wealthy	Safety	Lively	Beautiful	Boring	Depressing
Wealthy	1.000	0.954	0.978	-0.714	-0.143	0.848
Safety	0.954	1.000	0.950	-0.677	-0.284	0.874
Lively	0.978	0.950	1.000	-0.747	-0.192	0.884
Beautiful	-0.714	-0.677	-0.747	1.000	-0.203	-0.878
Boring	-0.143	-0.284	-0.192	-0.203	1.000	-0.204
Depressing	0.848	0.874	0.884	-0.878	-0.204	1.000

phenomenon was well demonstrated in urban studies of Chinese cities (Hanslmaier 2013, Song *et al*. 2018). In addition, the depression perception appears strongly related to wealthy, safety and lively perceptions. The relationship between urban development and the residents' depression status is a well-discussed issue in the field of public health and has gained increasing attention (Wang *et al*. 2018). These findings are consistent with those in the literature subserve the premise of the proposed methodology to identify local urban perceptions.

4.3. Spatial distribution of urban perceptions in the case study area

Based on the adversarial scoring embedded RF-model for fitting street-view perceptions, we calculated the average value of the four-direction street-view sampling points on different perceptions types. The resulting Wuhan urban perceptions distribution map is shown in Figure 8. Table 3 shows the statistical results of the perception scores of each administrative district in Wuhan. The old town area (Wuchang, Jiang'an, and Jianghan), which is a traditional commercial center and a densely populated area, obtained high wealthy, safety and lively perceptions with average scores that are approximately 1.0 to 1.2 times the overall level of Wuhan. Moreover, less boring perception scores were obtained in the old town area, outperforming the average level of Wuhan (≤ 60).

Dongxihu is a scenic spot district in the suburbs of Wuhan and has a relatively less-developed economy, thus receiving relatively low levels of wealthy, safety and lively perceptions while earning the highest beautiful and lowest boring and depressing perceptions within Wuhan. The derived local urban perception highly agreed with the economic development level of each administrative region, which demonstrated the efficacy of the proposed human-machine adversarial scoring framework for local urban perception assessment.

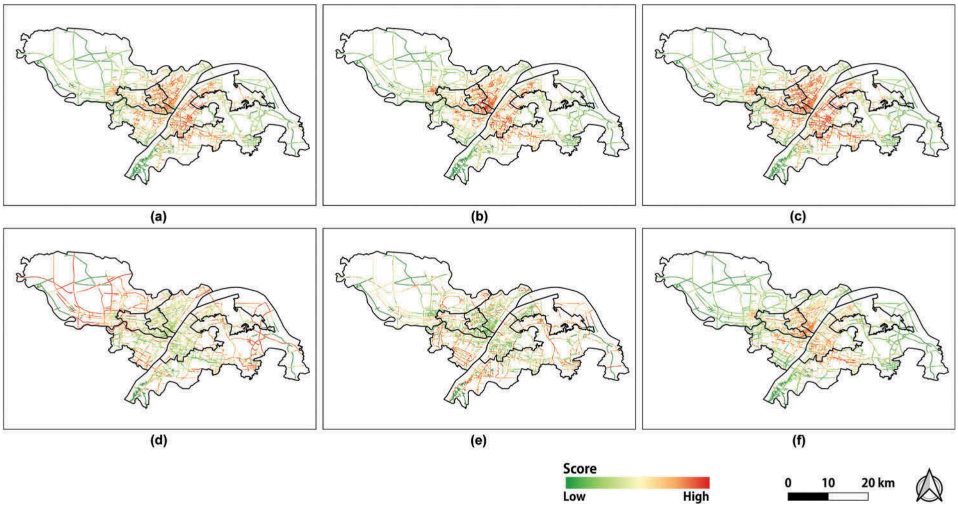

Figure 8. The distribution of urban perception results along the road. (a) Wealthy, (b) Safety, (c) Lively, (d) Beautiful, (e) Boring and (f) Depressing. (Low represents 0 score and high represent 100).

Table 3. Statistics of the urban perception results in the different administrative districts of the study area.

Perceptions	Statistic	Wuchang	Jianghan	Jiang'an	Qiaokou	Hanyang	Hongshan	Qingshan	Dongxihu	Wuhan
Wealthy	**Mean**	53.406	56.032	52.936	51.737	47.048	44.159	49.089	40.599	47.405
	Stdev.	6.511	2.645	6.468	5.825	6.451	8.567	6.181	7.174	8.655
	Median	54.822	56.300	55.264	53.222	48.080	44.755	50.207	39.800	49.240
Safety	**Mean**	47.533	52.145	48.294	46.737	40.635	38.519	43.743	36.374	42.148
	Stdev.	5.498	4.873	7.251	6.719	5.975	7.460	6.462	6.124	8.279
	Median	48.511	51.307	49.414	47.063	39.421	38.161	44.319	35.159	41.977
Lively	**Mean**	51.756	54.462	51.293	50.096	43.888	40.721	46.178	36.891	44.577
	Stdev.	6.116	3.035	7.050	6.752	7.429	10.361	7.810	8.959	10.211
	Median	52.877	54.756	54.009	52.250	44.877	41.616	47.160	36.168	47.250
Beautiful	**Mean**	38.236	36.534	39.169	40.072	41.074	43.458	43.212	50.619	42.812
	Stdev.	4.149	3.389	5.458	5.447	6.084	7.264	4.579	6.015	7.610
	Median	37.539	35.919	37.888	39.173	41.014	43.571	42.273	50.739	41.753
Boring	**Mean**	60.140	58.364	59.271	59.872	63.241	62.571	60.331	59.085	61.034
	Stdev.	2.568	2.829	3.078	2.716	2.563	3.078	3.687	2.736	3.331
	Median	59.816	59.011	59.097	59.876	63.683	62.413	59.415	60.119	60.905
Depressing	**Mean**	60.799	63.912	60.254	58.945	55.022	52.782	54.814	48.165	55.045
	Stdev.	4.462	4.012	6.684	6.414	7.011	8.228	5.719	5.480	8.213
	Median	61.674	64.734	61.305	59.112	54.044	51.272	56.834	47.261	55.015

4.4. Driving factor identification for the derived urban perception

4.4.1. Visual factor identification

Based on the parametric weight analysis of the random forest model (Palczewska *et al.* 2014), we analyzed the effects of 151 ground objects of 6 urban perception types and quantitatively investigated their interrelationships. As shown in Figure 9, the impact of the building layout contributes 15% to 30% to all urban perceptions except boring. Natural landscapes, including the sky, trees, rivers and grass, considerably impact residents' perceptions, especially in the depressing, boring and beautiful categories. Additionally, inspired by the close correlation between urban perceptions and urban land use (e.g., houses, fountains, and dirt tracks) revealed by the diagrams, we were interested in the potential relationship between urban perceptions and urban functional structures since urban land use always comes with a typical urban function.

4.4.2. Poi-based urban functional factor identification

To further study the relationship between urban functional patterns and urban perceptions, this study constructed nonlinear RF models to fit the urban perceptions and POI spatial distribution in the case study area. The case study included nine types of Gaode POIs, and used the mean integrated squared error (MISE) criterion (Duong and Hazelton 2003) to automatically determine the bandwidth of the Gaussian function-based kernel density analysis of POIs.

The fitting accuracy between the POIs and the urban perception distribution is shown in Table 4. All urban perceptions show a good fitting accuracy (R2 > 0.7, Person R > 0.85), but the perception of boring appears relatively weak compared to other perceptions. Our result shows that boring was a more subjective perception with a large deviation, which was consistent with the findings of (Zhang *et al.* 2018c).

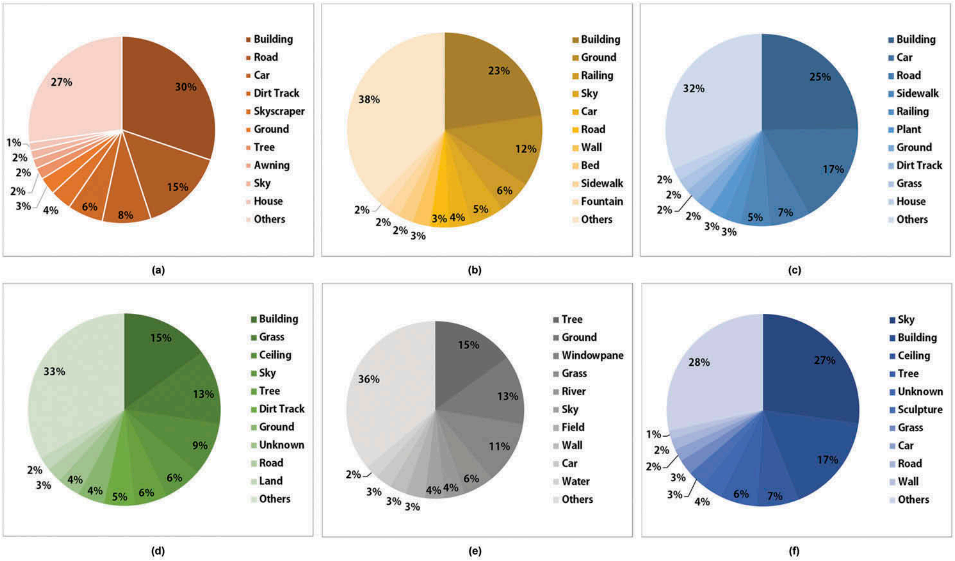

Figure 9. The weight of each ground object's impact on the different urban perceptions: (a) Wealthy, (b) Safety, (c) Lively, (d) Beautiful, (e) Boring and (f) Depressing.

Table 4. Fitting accuracy of the urban perceptions based on the POI densities via RF.

Perceptions	R^2	Pearson R	RMSE	MAE	OOB RMSE	OOB MAE
Wealthy	0.893	0.945	5.367	3.039	5.406	3.099
Safety	0.908	0.953	4.360	2.475	4.297	2.496
Lively	0.907	0.952	5.499	3.122	5.543	3.176
Beautiful	0.879	0.937	5.166	2.989	5.146	3.017
Boring	0.744	0.863	5.786	3.345	5.645	3.362
Depressing	0.903	0.951	5.394	2.956	5.204	2.914

Table 5. The fitting weights of the RF between the urban perceptions and POI categories: residential communities (RES), traffic facilities (TRA), commercial and business (COM), tourist attractions (TOU), food and shopping (SHP), education facilities (EDU), government and public services (GOV), financial services (FIN), and public facilities (PUB).

Perceptions	RES	TRA	COM	TOU	SHP	EDU	GOV	FIN	PUB
Wealthy	0.083	0.088	0.071	0.063	0.121	0.172	0.151	0.101	0.151
Safety	0.060	0.134	0.070	0.061	0.102	0.184	0.173	0.108	0.110
Lively	0.072	0.114	0.070	0.059	0.121	0.177	0.150	0.114	0.125
Beautiful	0.072	0.080	0.070	0.090	0.144	0.204	0.166	0.082	0.093
Boring	0.089	0.092	0.109	0.089	0.125	0.139	0.141	0.146	0.070
Depressing	0.057	0.084	0.073	0.081	0.092	0.239	0.228	0.064	0.083

The fitting weights between urban perceptions and POI categories are shown in Table 5. Education, government and shopping appear the most important factors influencing the urban perception in the case study area. In addition, the public facilities function has a greater weight for wealth perception compared with its weights for the others, which is consistent with common sense that developed and sufficient public facilities are commonly located in a wealthy area. The perceptions of safety and lively have very strong relationships with the traffic function. The existence of traffic facilities significantly contribute to safety, while the traffic flow volumes that are affected by the traffic function can substantially influence human impressions of the urban vitality.

5. Discussion and conclusion

Based on street-view data, this study proposed a rapid and cost-effective methodology for local urban perception assessment. This study used the proposed framework to estimate the urban and regional perceptions with high accuracy (RMSE within 10%), thereby indicating that human-machine adversarial scoring is useful for assisting local urban perception assessment. Experiments on identifying the driving factors demonstrated the feasibility and efficacy of the proposed framework.

This study proposed the idea of human-machine adversarial scoring to assess city-scale local urban perception in a cost-effective and accurate way. In the past, collection methods for city-scale human perceptions were limited to traditional interviews and questionnaire methods, which were labor-insensitive and time-consuming. Recent online crowd-sourcing strategies could improve this situation, but they still suffered from two shortcomings: (1) high time costs and economic costs for acquiring sufficient volunteers and (2) spatial heterogeneity for different countries and regions. Questions remained whether applications of datasets obtained from the information that was designed specifically for certain regions would be effective in other areas. The proposed human-machine adversarial scoring methodology addressed these questions and provided a rapid and cost-effective way to accurately generate city-scale data on local urban perceptions.

We designed a case study to demonstrate the feasibility of the proposed methodology. The number of volunteers recruited in the case study was small (only 20 citizens). In order to solve the problem that the small number of volunteers may lead to data bias, we are developing a website for the human-machine adversarial-based perception scoring system that supports crowdsourcing for future study. We will build a specific sensory dataset suitable for supporting large cities and regions in China and provide available urban perception data and services for urban planners.

By exploring the correlation and fitting weights between the urban perceptions and urban functional patterns revealed by the POIs, we found a very strong nonlinear relationship between urban perceptions and urban functional structures. POIs were proven to be useful for urban landscape evaluations (Liu *et al*. 2017, Waddell *et al*. 2010, Yao *et al*. 2016), and this study also found that the POIs could accurately estimate the distribution of urban perceptions (the RMSE approximated 5%). The case study quantitatively identified the impact of urban functions on urban perceptions and demonstrated the efficacy of the proposed adversarial scoring methodology in facilitating local urban perception assessment.

The proposed method has many limitations and opportunities for future studies. Urban perceptions are unique and subjective, not only related to the street scenery seen by the individual but also to other factors in the city, such as the noise, temperature, humidity, commodity prices, etc. (Bonaiutoa *et al*. 1999, Hong and Jin 2015, Gunnarsson *et al*. 2017). Therefore, future studies need to consider more factors in the design of urban perception models with diverse input data (such as street scenes, videos, etc.), and evaluation goals (such as noise, livability, etc.) to develop a more complete and accurate urban perception analysis model. In addition, when identifying urban functional driving factors, the weights obtained by the RF model can only indicate the importance of the variables to the results but not decide whether the driving factors are positive or negative (Biau 2012). Besides, the collection speed of street-view imagery may not keep apace with the change of urban landscape and the mapping result of a city's urban perception may have temporal issue. We will focus on these issues in future research.

This study did not consider the issue of ethnic differences in volunteers. According to 2010 Census data of National Bureau of Statistics of China (2011) in China, the proportion of minority in China is 8.49%, while the proportion of minority in Wuhan is 0.90%, so the potential bias caused by the existing of some particular ethnic group may be negligible in our results. The volunteers in our study are from both northern and southern cities (which may be different in cultural and other aspects), so their rating behaviors may be relatively representative and can weaken the effect of genetic or cultural homogeneity in China. However, the future study may be restricted to a particular ethnic group and thus ensure better genetic or cultural homogeneity. Moreover, we will record as much personal information as possible from volunteers in the future study. Therefore, we can further investigate biased effect of the scenes which are scored and operated differently by either one or multiple volunteers. The relationship between gender, age, income level, ethnicity and the perceptions of the urban environment is expected to be better informed by considering such biases from annotators.

Further studies should consider whether the 'recommendation score' would bias the choice of volunteers' respondents. On the one hand, we consider that this technique is able to give a reasonable recommendation score without bias in most cases because we assume that respondents, after 50 images, should have a conditioned reflex on image scoring. The human-machine adversarial program is designed to capture such a conditioned reflex on image scoring that subconsciously leads respondents to score the next image. This kind of phenomenon has been proved by previous studies. For example, works done by Anokhin (2016) and Lang (2000) indicated that after several repeated actions, respondents may act without hesitation, which would be much faster than before. The recommendation score might not seriously bias respondents' actions. Jones *et al*. (2009) used a similar method to score the cell morphologies in an image via machine learning and iterative feedback and improved the phenotype identification efficiency.

On the other hand, this technique may lead to bias for some respondents because of the perceptual prime effect. Respondents' perception score on a middle-income neighborhood may be biased under different imagery displaying strategy, e.g., a series of images of wealthy neighborhoods vs. a series of images of poor neighborhoods. Future studies may revise the method, such as sorting the images according to existing residential segregation phenomenon. A revised method may sort images into various neighborhood contexts in advance and make batches of display images highly consistent. The revised method can keep records of image displaying orders and examine the potential bias caused by the perceptual prime effect.

Urban perceptions can be very important to the field of public health (Hong and Jin 2015, Wang *et al*. 2018) and can also be integrated into the best urban planning practices by considering the feelings of local urban residents. To cost-effectively assess urban perceptions at a local scale, this study first proposed an effective human-machine adversarial scoring framework that incorporates deep learning using street-view imagery Perception data from the research can provide a basis for spatial correlation mining between public health data and local urban perception on street scenery.

Additionally, this study showed a strong correlation between urban functional patterns and urban perceptions and quantitatively identified driving factors of urban features for different urban perceptions. The study demonstrated the efficacy of the proposed adversarial scoring methodology in facilitating local urban perception assessments. By taking advantage of the enriched spatial semantics using human perceptions, the proposed framework is able to help researchers understand the underlying urban structure and reveal the impacts of urban function using an affordable and rapid solution, thereby facilitating urban planners in integrating urban perceptions into their planning practices for more sustainable and human-oriented urban development.

Notes

1. The ADE-20K is an open data set that can be downloaded from the MIT website (http://groups.csail.mit.edu/vision/datasets/ADE20K/).
2. Official website of the MIT ADE20K dataset: http://groups.csail.mit.edu/vision/datasets/ADE20K/.
3. The 8 administrative districts are as follows: Wuchang, Hongshan, Jiang'an, Qiaokou, Hanyang, Jianghan, Qingshan and Dongxihu. The residential population of each district is 1,178 million, 1,107 million, 755 million, 723 million, 673 million, 661 million, 502 million and 374 million, respectively.

Acknowledgments

We fully appreciate Prof. May Yuan, Prof. Robert Weibel and the anonymous reviewers for their helpful comments and suggestions. We are also very grateful to the volunteers who helped us with this work. This study was supported by the National Key R&D Program of China [Grant No. 2017YFB0503800]; the National Natural Science Foundation of China [Grant No. 41801306, 41671408]; the Open Fund of State Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University [Grant No. 18S01] and by the Fundamental Research Founds for National University, China University of Geosciences (Wuhan) [Grant No. CUG190606].

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [41671408,41801306]; Fundamental Research Founds for National University, China University of Geosciences(Wuhan) [CUG190606]; Open Fund of State Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,Wuhan University [18S01]; National Key R&D Program of China [2017YFB0503804].

Notes on contributors

Yao Yao is an Associate Professor at China University of Geosciences (Wuhan) and Senior Algorithm Engineer at Alibaba Group. His research interest is spatio-temporal big data mining and urban comupting.

Zhaotang Liang obtained his master's degree from the Chinese University of Hong Kong. Now he is a research assistant at the Chinese University of Hong Kong. His research interest is geospatial big data analysis.

Zehao Yuan is a master student at China University of Geosciences (Wuhan). His research interest is urban function analysis using social media data.

Penghua Liu is a master student at Sun Yat-sen University. His research interest is geospatial data analysis and modeling.

Yongpan Bie is a master student at China University of Geosciences (Wuhan). His research interest is Spatial data visualization and analysis.

Jinbao Zhang is a PhD. candidate at Sun Yat-sen University and Visiting scholar at Tencent Technology Inc. His research interest is urban computing and modeling using social media data.

Ruoyu Wang is a PhD. candidate at University of Edinburgh and research assistant at Sun Yat-sen University. His research interest is health geography.

Jiale Wang is a master student at Wuhan University. His major is cartography and geographic information systems.

Qingfeng Guan is a Professor at China University of Geosciences (Wuhan). His research interest is high-performance spatial intelligence computing and spatio-temporal data mining.

ORCID

Yao Yao http://orcid.org/0000-0002-2830-0377
Zhaotang Liang http://orcid.org/0000-0001-9261-5261
Penghua Liu http://orcid.org/0000-0002-8574-891X
Jinbao Zhang http://orcid.org/0000-0001-8510-149X
Qingfeng Guan http://orcid.org/0000-0002-7392-3709

References

Anguelov, D., *et al.*, 2010. Google street view: capturing the world at street level. *Computer*, 43 (6), 32–38. doi:10.1109/MC.2010.170

Anokhin, P.K., 2016. *Biology and neurophysiology of the conditioned reflex and its role in adaptive behavior. In*: *International Series of Monographs in Cerebrovisceral and Behavioral* Physiology *and* Conditioned Reflexes, Elsevier.

Ashihara, Y., 1983. *The aesthetic townscape*. Cambridge: MIT Press.

Badrinarayanan, V., Kendall, A., and Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 99, 2481–2495. doi:10.1109/TPAMI.2016.2644615

Biau, G.E.R., 2012. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13 (1), 1063–1095.

Bonaiutoa, M., Aielloa, A., and Peruginib, M., 1999. Multidimensional perception of residential environment quality and neighborhood attachment in the urban environment. *Journal of Environmental Psychology*, 19 (4), 331–352. doi:10.1006/jevp.1999.0138

Cameron, L. and Larsen-Freeman, D., 2007. Complex systems and applied linguistics. *International Journal of Applied Linguistics*, 17 (2), 226–239. doi:10.1111/j.1473-4192.2007.00148.x

Cheng, L., *et al.*, 2017. Use of tencent street view imagery for visual perception of streets. *International Journal of Geo-Information*, 6 (9), 265. doi:10.3390/ijgi6090265

Dadvand, P., *et al.*, 2016. Green spaces and general health: roles of mental health status, social support, and physical activity. *Environment International*, 91, 161–167. doi:10.1016/j.envint.2016.02.029

Dubey, A., *et al.*, 2016. *Deep learning the city: quantifying urban perception at a global scale*. Cham: Springer International Publishing, 196–212.

Duong, T. and Hazelton, M., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15 (1), 17–30. doi:10.1080/10485250306039

Fern A Ndez-Delgado, M., *et al.*, 2014. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15 (1), 3133–3181.

Frank, L.D. and Engelke, P.O., 2001. The built environment and human activity patterns: exploring the impacts of urban form on public health. *Journal of Planning Literature*, 16 (2), 202–218. doi:10.1177/08854120122093339

Gebru, T., *et al*., 2017. Using deep learning and google street view to estimate the demographic makeup of the US. *Proceedings of the National Academy of Sciences*, 114 (50), 13108–13113. doi:10.1073/pnas.1700035114

Greene, M.R. and Oliva, A., 2009a. Recognition of natural scenes from global properties: seeing the forest without representing the trees. *Cognitive Psychology*, 58 (2), 137–176. doi:10.1016/j.cogpsych.2008.06.001

Greene, M.R. and Oliva, A., 2009b. The briefest of glances. *Psychological Science*, 20 (4), 464–472. doi:10.1111/j.1467-9280.2009.02316.x

Gunnarsson, B., *et al*., 2017. Effects of biodiversity and environment-related attitude on perception of urban green space. *Urban Ecosystems*, 20 (1), 37–49. doi:10.1007/s11252-016-0581-x

Halpern, D., 1995. *Mental health and the built environment: more than bricks and mortar?* Taylor & Francis Ltd.

Hannay, D.R., 1983. Mental illness in the community: the pathway to psychiatric care. *International Journal of Rehabilitation Research*, 6 (s37), 47–53.

Hanslmaier, M., 2013. Crime, fear and subjective well-being: how victimization and street crime affect fear and life satisfaction. *European Journal of Criminology*, 10 (5), 515–533. doi:10.1177/1477370812474545

Hong, J.Y. and Jin, Y.J., 2015. Influence of urban contexts on soundscape perceptions: A structural equation modeling approach. *Landscape & Urban Planning*, 141, 78–87. doi:10.1016/j.landurbplan.2015.05.004

Hu, Y., *et al*., 2015. Extracting and understanding urban areas of interest using geotagged photos. *Computers Environment & Urban Systems*, 54, 240–254. doi:10.1016/j.compenvurbsys.2015.09.001

Jones, T.R., *et al*., 2009. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (6), 1826–1831. doi:10.1073/pnas.0808843106

Kabisch, N., Qureshi, S., and Haase, D., 2015. Human-environment interactions in urban green spaces - A systematic review of contemporary issues and prospects for future research. *Environmental Impact Assessment Review*, 50 (50), 25–34. doi:10.1016/j.eiar.2014.08.007

Lang, P.J., 2000. Emotion and motivation: attention, perception, and action. *Journal of Sport and Exercise Psychology*, 22 (S1), S122–S140. doi:10.1123/jsep.22.s1.s122

Liu, X., *et al*., 2017. Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science*, 31 (8), 1675–1696. doi:10.1080/13658816.2017.1324976

Long, J., Shelhamer, E., and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Long, Y. and Liu, L., 2017. How green are the streets? An analysis for central areas of Chinese cities using tencent street view. *PloS One*, 12 (2), e171110. doi:10.1371/journal.pone.0171110

Naik, N., *et al*., 2017. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114 (29), 7571–7576. doi:10.1073/pnas.1619003114

Ordonez, V. and Berg, T.L., 2014. Learning high-level judgments of urban perception. *In*: *European conference on computer vision*, Cham: Springer, 494–510. doi:10.1016/j.msec.2014.06.046

Palczewska, A., *et al*., 2014. Interpreting random forest classification models using a feature contribution method. *In*: *Integration of reusable systems*. Cham: Springer, 193–218.

Porzi, L., *et al*., 2015. Predicting and understanding urban perception with convolutional neural networks. 139–148.

Rapoport, A. and Hawkes, R., 1970. The perception of urban complexity. *Journal of the American Institute of Planners*, 36 (2), 106–111. doi:10.1080/01944367008977291

Salesses, P., Schechtner, K., and Hidalgo, C.E.S.A., 2013. The collaborative image of the city: mapping the inequality of urban perception. *PloS One*, 8 (7), e68400. doi:10.1371/journal.pone.0068400

Shelhamer, E., Long, J., and Darrell, T., 2014. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39 (4), 1.

Song, G., *et al.*, 2018. Testing indicators of risk populations for theft from the person across space and time: the significance of mobility and outdoor activity. *Annals of the American Association of Geographers*, 108 (5), 1370–1388. doi:10.1080/24694452.2017.1414580

Sun, A., Chen, T., and Niu, R., 2016. Urbanization analysis in Wuhan area from 1991 to 2013 based on MESMA. *In*: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Beijing, 5473–5476.

Tuan, Y.F., 2013. Landscapes of fear. University of Minnesota Press.

Ulrich, R.S., 1979. Visual landscapes and psychological well-being. *Landscape Research*, 4 (1), 17–23. doi:10.1080/01426397908705892

Waddell, P., *et al.*, 2010. Microsimulating parcel-level land use and activity-based travel: development of a prototype application in San Francisco. *Journal of Transport and Land Use*, 3, 2. doi:10.5198/jtlu.v3i2.124

Wang, R., *et al.*, 2018. The relationship between urbanization and depression in China: the mediating role of neighborhood social capital. *International Journal for Equity in Health*, 17 (1), 105. doi:10.1186/s12939-018-0825-x

Wolch, J.R., Byrne, J., and Newell, J.P., 2014. Urban green space, public health, and environmental justice: the challenge of making cities "just green enough". *Landscape & Urban Planning*, 125, 234–244. doi:10.1016/j.landurbplan.2014.01.017

Wong, H.W., 2015. *Birds in a cage: political institutions and civil society in Hong Kong*. Electoral Politics in Post-1997 Hong Kong. Singapore: Springer, 45–68.

Xu, G., *et al*. 2017. Automatic land cover classification of geo-tagged field photos by deep learning. *Environmental Modelling & Software*, 91, 127–134.

Yao, X., Wang, Z., and Zhang, H., 2016. Dynamic changes of the ecological footprint and its component analysis response to land use in Wuhan, China. *Sustainability*, 8 (4), 329. doi:10.3390/su8040329

Yao, Y., *et al*., 2016. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science*, 31 (4), 825–848.

Yao, Y., *et al*., 2017a. Simulating urban land-use changes at a large scale by integrating dynamic land parcel subdivision and vector-based cellular automata. *International Journal of Geographical Information Science*, 31 (12), 2452–2479. doi:10.1080/13658816.2017.1360494

Yao, Y., *et al*., 2017b. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science*, 31 (6), 1220–1244.

Yao, Y., *et al*., 2018. Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Transactions in GIS*, 22 (2), 561–581. doi:10.1111/tgis.2018.22.issue-2

Zhang, F., *et al*., 2018a. Framework for virtual cognitive experiment in virtual geographic environments. *International Journal of Geo-Information*, 7 (2), 36.

Zhang, F., *et al*. 2018b. Representing place locales using scene elements. *Computers Environment & Urban Systems*, 71, 153–164.

Zhang, F., *et al*. 2018c. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180, 148–160.

Zheng, S., *et al*., 2016. Conditional random fields as recurrent neural networks. *In*: *Proceedings of the IEEE international conference on computer vision*, Santiago, 1529–1537.

Zhou, B., *et al*., 2014. Recognizing city identity via attribute analysis of geo-tagged images. *In*: *European conference on computer vision*. Cham: Springer, 519–534.

Zhou, B., *et al*., 2016. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127 (3), 302–321.

Zhou, B., *et al*., 2017. Scene Parsing through ADE20K Dataset. *In*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, 633–641.

Zhou, B., *et al*., 2019. Semantic understanding of scenes through the ADE20k dataset. *International Journal of Computer Vision*, 127 (3), 302–321.