Wang *et al. Computational Urban Science* (2021) 1:3
https://doi.org/10.1007/s43762-021-00003-0

Computational Urban Science

ORIGINAL PAPER Open Access

A comparison of two deep-learning-based urban perception models: which one is better?

Ruifan Wang[1†], Shuliang Ren[1†], Jiaqi Zhang[1†], Yao Yao[1,2*], Yu Wang[1] and Qingfeng Guan[1]

Abstract

Urban perception is a hot topic in current urban study and plays a positive role in urban planning and design. At present, there are two methods to calculate urban perception. 1) Using a model to learn image features directly automatically; 2) Coupling machine learning and feature extraction based on expert knowledge (e.g. object proportion) method. With two typical streets in Wuhan as the study area, video data were recorded and used as the model input. In this study, two representative methods are selected: 1) End to end convolution neural network (CNN-based model); 2) Based on full convolution neural network and random forest (FCN + RF-based model). By comparing the accuracy of two models, we analyze the adaptability of the model in different urban scenes. We also analyze the relationship between CNN-based model and urban function based on POI data and OSM data, and verify its interpretability. The results show that the CNN-based model is more accurate than FCN + RF-based model. Because the CNN-based model considers the topological characteristics of the ground objects, its perception results have a stronger nonlinear correlation with urban functions. In addition, we also find that the CNN-based model is more suitable for scenes with weak spatial heterogeneity (such as small and medium-sized urban environments), while the FCN + RF-based model is applicable to scenes with strong spatial heterogeneity (such as the downtown areas of China's megacities). The results of this study can be used as a reference to provide decision support for urban perception model selection in urban planning.

Keywords: Urban perception, Deep learning, Urban functions, Street views, Random forest

1 Introduction

Urban perception refers to people's feelings about the urban visual environment, that is, their esthetic judgment of the urban scene (Weber, Schnier, & Jacobsen, 2008). In recent years, urban perception studies have developed rapidly. The results have been applied to many aspects of urban construction, such as urban space esthetics, urban safety, and urban public health (Harvey et al., 2015; Helbich et al., 2019; Weber et al., 2008). The Chinese government puts forward the idea of people-oriented urban planning (Blaxland, Shang, & Fisher, 2014). People's perceptions of the city are of great value in future urban planning and design (Been et al., 2016; Cheng et al., 2017; Ozkan, 2014). Therefore, it is very important to study the method of urban perception for decision support in urban management, urban planning and policymaking.

Traditional perception data are collected by social scientists through field investigation (Liu et al., 2015; Sampson, 2012). This method is time-consuming, expensive and has a small study scope. Street view objectively depicts the real urban landscape, which has been proven to be effective and reliable data to measure the urban environment (Long & Liu, 2017; Zhang et al., 2018a, b, c). In recent years, perception prediction based

* Correspondence: yaoy@cug.edu.cn
†Ruifan Wang, Shuliang Ren and Jiaqi Zhang contributed equally to this work.
[1]School of Geography and Information Engineering, China University of Geoscience, Wuhan 430078, Hubei, China
[2]Alibaba Group, Hangzhou 311121, Zhejiang, China

on street view images has received increasing attention. Salesses, Schechtner, and Hidalgo (2013) found that street view images can be used to assess the social and economic impact of the urban environment. Naik et al. (2014) proposed a method of scene perception prediction based on support vector regression. However, this method has the problem of relying on predefined feature mapping. With the popularity of deep learning, most scholars have begun to introduce it into the study of urban perception (Dubey et al., 2016; Liu et al., 2017a, b). Porzi et al. (2015) proved that deep learning is superior to traditional feature description in predicting human perception. Naik, Raskar, and Hidalgo (2016) and Zhang et al. (2018a, b, c) used computer vision to simulate individuals to quantify urban perception scores.

At present, the dataset used in urban perception is mainly the place pulse dataset provided by MIT (Ordonez & Berg, 2014; Porzi et al., 2015). This urban perception dataset has collected street view images from many cities around the world but lacks training samples for the cities of mainland China. There are obvious differences in architectural styles and town planning between the East and the West (Ashihara, 1983). The urban perception model derived from this global dataset may not be applicable to the cities in mainland China. Therefore, Yao et al. (2019) proposed building a unique urban perception dataset in China. They constructed a human-machine adversarial scoring framework based on deep learning to support the perception study of cities and regions in mainland China and finally obtained interpretable results.

The above work shows that urban perception based on street view images and deep learning technology is the main study direction, which can deepen people's understanding of large-scale urban environment scenes in a more automatic and effective way. There are two ways to obtain urban perception score through deep learning. The first is to learn the deep features automatically via the deep learning model and fit the urban perception score. For example, Dubey et al. (2016) used an end-to-end model to directly extract the high-dimensional features of street view images to predict the urban perception scores. Zhang et al. (2018a, b, c) used the street view images as the input, acquired the scores as the output and predicted the perception scores of the images in six dimensions.

Another method is to obtain the features (object proportion) constructed by expert knowledge based on scene semantic deep learning model and then fit them by machine learning. For example, the FCN + RF based model proposed by Yao et al. (2019) firstly uses the full convolution neural network to identify the ground objects and then obtains urban perception score based on the proportion of ground objects and random forest. This method has been used in the field of public health (Wang et al., 2019a, b), and the results are reasonable and interpretable. Compared with the black box mechanism of the CNN perception model, the perception model of FCN + RF-based model proposed by Yao et al. (2019) obtains artificial features from expert knowledge and has better interpretability.

The two deep learning methods have different principles in obtaining features and fitting urban perception scores. This leads to a discussion of the differences between the two models. Therefore, this study proposes two issues for further discussion: 1) What are the advantages and disadvantages of the two methods in the urban perception prediction task? 2) Is the result of automatic feature extraction based on the deep learning model reasonable explanation? At present, there is no relevant study on this issue.

To solve the above problems, this study constructs a typical model (CNN-based model) which can automatically learn deep features and compares its score results with FCN + RF-based model (Yao et al., 2019). We train the CNN-based model based on the China urban perception dataset (Yao et al., 2019). Then, we use the street view images that we collected by video to explore the differences in perception scores between the two methods and analyze models' scene suitability. To verify the interpretability of the CNN-based model, we use point of interests (POI) and OpenStreetMap (OSM) data to explore the drivers that affect different perceptions.

2 Methodology

Figure 1 displays the flow chart of this study. First, we develop a mobile app that can capture driving scenes and obtain real-time longitude and latitude data at the same time. According to app's time-series data, we process the mobile video files to get the image dataset of the study area. Second, we use the China urban perception dataset (Yao et al., 2019) as the training dataset to train six perception models (beautiful, wealthy, depressing, lively, safety, boring) and quantify the street view images of the study area. Third, based on the street view images collected while driving we evaluate the accuracy of the model results and analyze the scene suitability of the CNN-based model and FCN + RF-based model. Then, POI and OSM road network data are used to analyze the driving factors of different urban perceptions, and the interpretability of CNN-based model is verified.

2.1 Mobile video access to street view images

According to the time series data collected by GPS, the adjacent points' timestamps are calculated. The study area's street view images are extracted from the video file by using the time stamp information. The code is

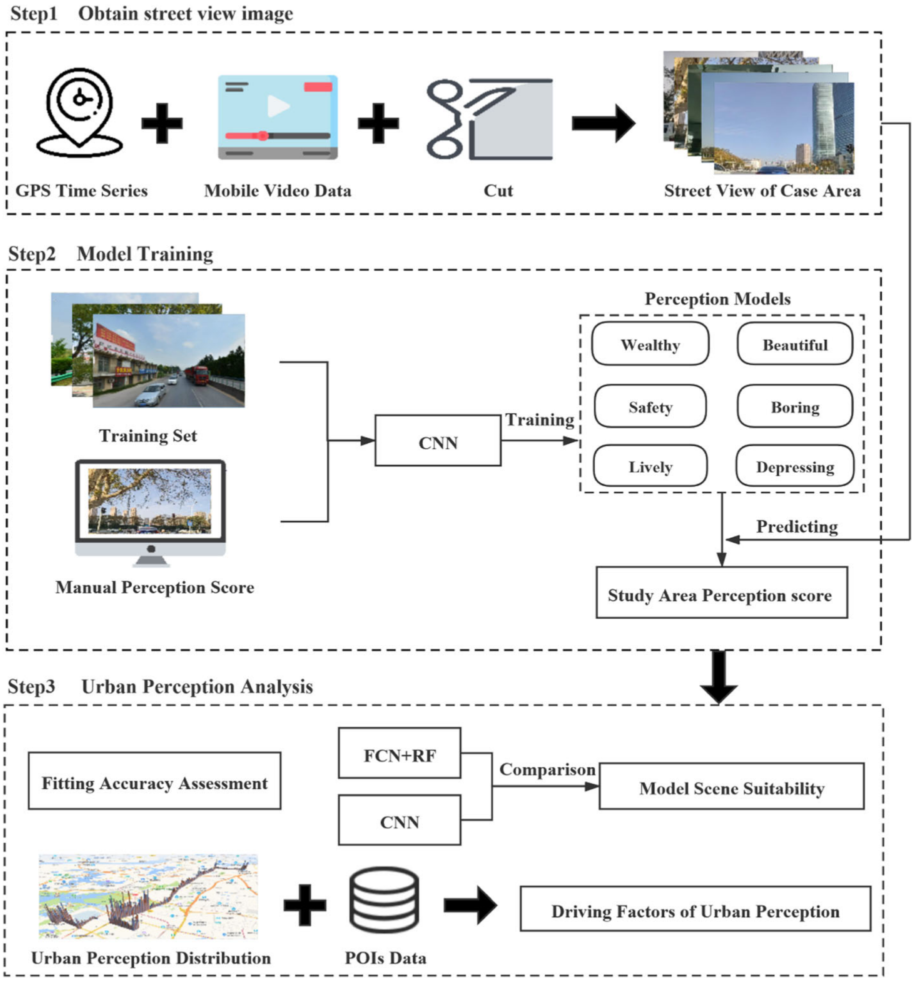

Fig. 1 Workflow of assessing street-level local urban perception

freely available at https://github.com/Leitast/LocationListener. In this study, a mobile phone is placed in a vehicle for video recording. The bottom of the scene, which includes the vehicle, is cut to ensure the experiment's reliability.

2.2 CNN-based urban perception model

The depth of the network is critical to the performance of the model. When the number of network layers is increased, the network can extract more complex features (Simonyan & Zisserman, 2014). To extract enough image features, we construct an end-to-end CNN model to represent urban perception. The end-to-end perception model proposed in this study refers to the structure of VGGNet (Simonyan & Zisserman, 2014). The model's parameters and structure have been proved to be reliable in many aspects of image feature extraction (Ha et al., 2018; Liu et al., 2018; Lu et al., 2017). Increasing network's depth may lead to degradation, and may cause decrease of testing and training accuracy results (Monti, Tootoonian, & Cao, 2018). Considering the small number of training samples in this study, the network depth is reduced to avoid degradation. By repeatedly stacking a 3 × 3 convolution kernel and a 2 × 2 maximum pooling layer, a model for urban perception is constructed. In this study, batch normalization is used to speed up the convergence of the model and avoid gradient dispersion. Batch normalization is better than dropout (Ioffe & Szegedy, 2015).

Traditional CNN models perform well in image classification (He et al., 2016; Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2014). In the field of urban study, CNNs have been used for the semantic segmentation of urban traffic scenes and land-use change analysis (Deng et al., 2017; Zhai et al., 2020). As shown in Fig. 2, to preserve the image feature structure, we replace the softmax, which is responsible for multiclassification tasks, with a one-dimensional full connection layer to realize the end-to-end function. By inputting a street view image, the model can extract the topological

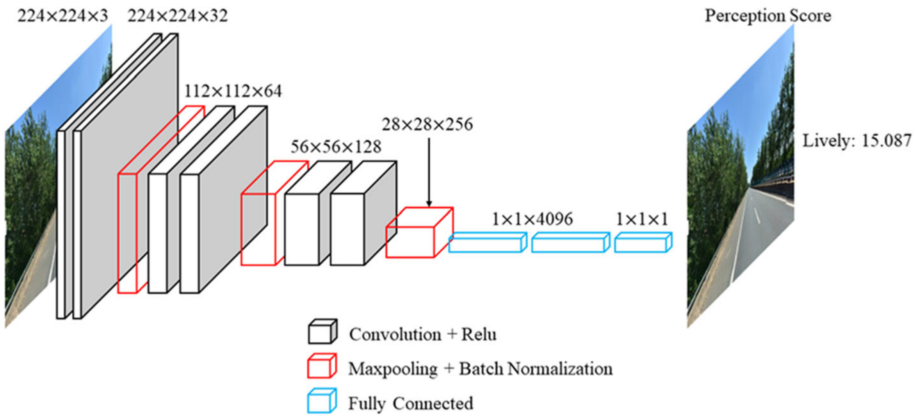

Fig. 2 The computational framework of the CNN model

features of the image and obtain the urban perception score.

2.3 FCN + RF-based urban perception model

The FCN + RF-based model (Yao et al., 2019) is used in this study for comparative analysis. FCN can predict the semantic features of each pixel in the image to generate natural target level segmentation results and obtain the classification of each image (such as sky, road, car, and building) (Badrinarayanan, Kendall, & Cipolla, 2017; Cordts et al., 2016; Long, Shelhamer, & Darrell, 2015). Yao et al. (2019) selected the annotated images in the ADE-20 k scene analysis and segmentation database (Zhou et al., 2017; Zhou et al., 2019). Taking the street view images as the input of the FCN, the ratio of 151 categories in each image is obtained. Finally, the ratios of the 151 categories are used as the input of RF to obtain the scores of each urban perception. The model has been applied in public health (Wang et al., 2019a, b) and proved to be reliable and effective. The detailed structure of the model can refer to the paper of Yao et al. (2019). The software can be found at http://www.urbancomp.net/2020/08/03/semantic-segmentation-software-for-visual-images-based-on-fcn.

2.4 Interpretability analysis of urban perception based on POI and OSM data

POI data have been applied to urban functional area identification (Yuan, Zheng, & Xie, 2012). Palczewska et al. (2014) demonstrated the correlation between urban function and urban perception using random forest. Yao et al. (2019) also verified the interpretability of the FCN + RF-based model from the perspective of urban function based on POI or OSM and Random Forest. Therefore, we get the urban perception distribution using the CNN-based model, analyze the correlation between the urban functional areas and the simulation results, and explore the model results' interpretability through the feature importance function of random forest.

2.5 Accuracy assessment

During the process of comparing the advantages and disadvantages of the different models and using POI and OSM to analyze the correlation between urban function and urban perception, this study uses the mean absolute error (MAE), root mean squared error (RMSE) and Pearson correlation coefficient (Pearson R) to quantify the accuracy between the predictions and the ground-truth values. The MAE, RMSE and Pearson R are mathematically represented by Eq. (1) to Eq. (3), respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right| \tag{1}$$

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

$$Pearson\ R = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)\left(\widehat{y_i} - \overline{\widehat{y_i}}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\widehat{y_i} - \overline{\widehat{y_i}}\right)^2}} \tag{3}$$

where y_i is the ground-truth value, $\overline{y}$ is equal to $\frac{1}{n}\sum_{i=1}^{n} y_i$, and $\widehat{y}_i$ is the predicted result.

3 Study area and data

As the largest political, economic and cultural center in Central China (Sun, Chen, & Niu, 2016), Wuhan has become one of the most rapidly developing cities in China. As shown in Fig. 3, this study selects two streets, LuMo Road and Gaoxin Avenue, which run through the suburbs and urban center of Wuhan. Figure 3(c) shows that the red line represents LuMo Road, connecting the school and the business center. The yellow line represents Gaoxin Avenue, which connects the satellite city

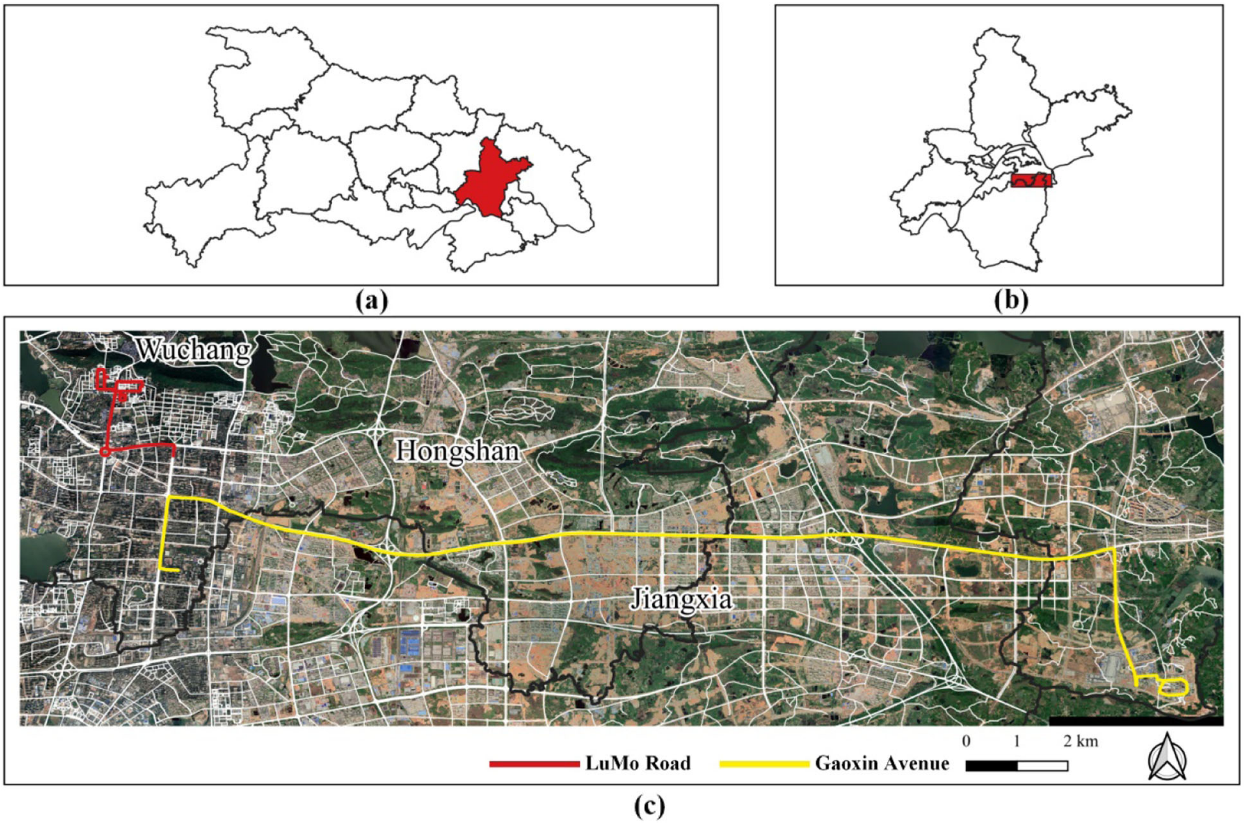

Fig. 3 The study area and its location in Hubei Province in China: **a** the geographic location of Wuhan City in Hubei; **b** the study area; **c** the red lines are LuMo Road and the yellow lines are Gaoxin Avenue. The white lines are the main roads in the study area obtained from openstreetmap.org

(Zuoling community) and the central city. From east to west, Gaoxin Avenue passes through undeveloped suburbs, satellite cities, science and technology industrial parks newly planned by the government, mature commercial district, and residential area. Therefore, the street view images collected from these two streets can cover various of urban functional areas in Wuhan. They have an excellent verification effect in the comparative analysis of the models.

We obtain the street view images of 3592 sample points for perception analysis. Among them, 1154 sample points are collected from LuMo Road, and 2798 sample points are collected from Gaoxin Avenue.

Figure 4 shows the street view images used in this study. (A), (B), (C) and (D) are the sample images of the China urban perception dataset (Yao et al., 2019) used for model training. Each image in the dataset is scored on six perceptions (wealthy, safety, lively, beautiful, boring and depressing, with a score range of 0–100 by volunteers who have a good understanding of the local socio-economic background). The geographic location of the images in the dataset is close to the study area. Therefore, this dataset is of great help for the perception analysis in the study area. (E) and (F) are street view images in the study area collected through mobile video.

POI and OSM data are also used in our study. Gaode, the largest online map service provider in China, has complete POI resources (http://amap.com). We obtain 24 categories of POI data in the study area (https://lbs.amap.com/api/webservice/download). To facilitate the follow-up study, we combine similar classes and calculate the kernel density (Fig. 5). The processed 12 kinds of POI and OSM road network data can describe and analyze the social economy and infrastructure (Liu et al., 2017a, b; Yao et al., 2018).

4 Results

4.1 Comparison of model accuracy based on street view dataset

In this study, the China urban perception dataset (Yao et al., 2019) is divided into a training set and a test set, where 80% of the data are used for training and 20% of the data are used for testing. We use the training set to build the fitting model and use the test set to evaluate the prediction results of the model. It should be noted that the parameters of the FCN module in FCN + RF-

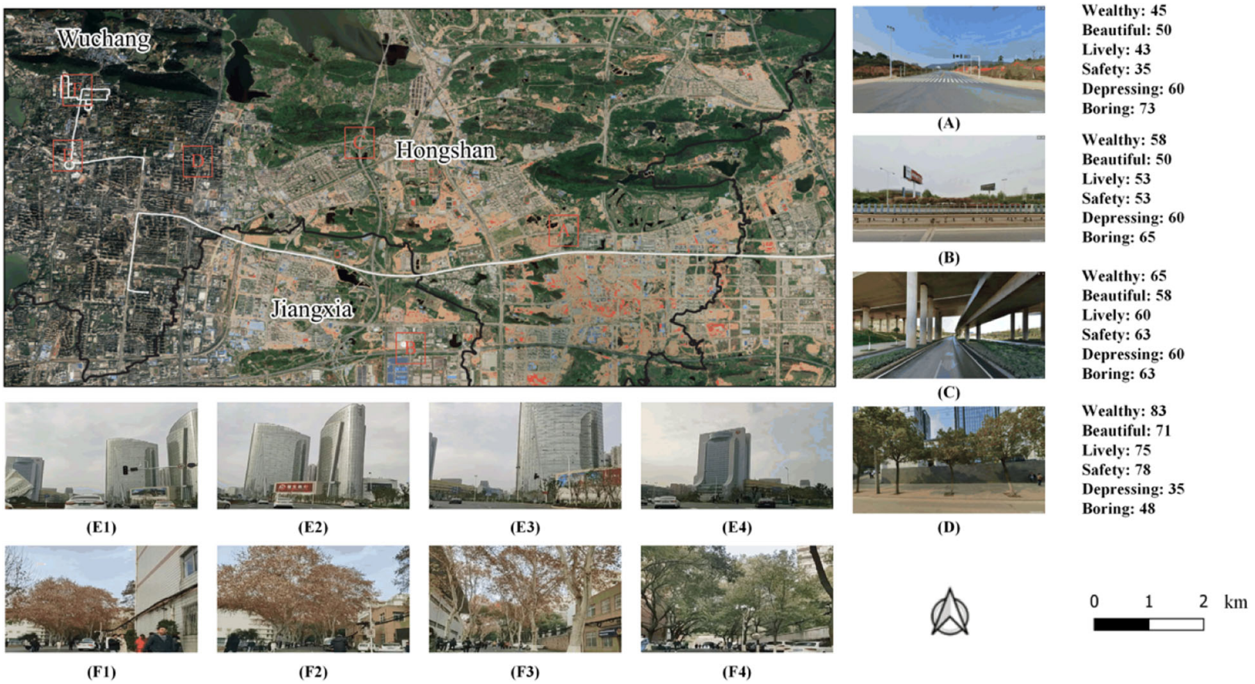

Fig. 4 Street view images involved in this study: (A) (B) (C) (D) are samples of street view images provided from the training data set. (E1- E4) and (F1- F4) are street view images obtained by a mobile phone. (E1)- (E4) are images of the Wuhan Optical Valley (CBD area), (F1)- (F4) are images of the campus of the China University of Geosciences

based model are set by referring to Long et al. (2015), while the RF part is formed by grid optimization. These models and parameters have been widely used in image semantic segmentation (Zhou et al., 2016), segmentation of street view images (Middel et al., 2019), remote sensing image classification (Han et al., 2020; Piramanayagam et al., 2016), public health (Zamani Joharestani et al., 2019), and proved to be effective. To better compare the CNN-based model, we refer to (Bulat & Tzimiropoulos, 2016; Simonyan & Zisserman, 2014) and other models to set parameters for the CNN's hyperparameters.

Table 1 shows the training accuracy results of the six perception models. The Pearson R of all perceptions is greater than 0.9, which shows that the CNN-based model proposed in this study has a good perception effect. The results of the FCN + RF-based model is shown in Table 2. We find that the fitting results based on CNN model (the average error of RMSE of each perception is around 6.5) are more accurate than those based on FCN + RF-based model (the average error of RMSE of each perception is approximately 9.1). The CNN can automatically extract the features from images (Sun, Li, & Huang, 2017). These features represent the color, contour, texture, and spatial structure of the objects in images (Hu et al., 2015; Jiao et al., 2017; Kim & Pavlovic, 2016; Sahiner et al., 1996). Compared with the method of calculating the proportion of ground objects, the CNN-based model can also learn highly representative and hierarchical image features from sufficient training data (Shin et al., 2016)

4.2 Comparison of model results based on real application environment

According to Tables 1 and 2, the accuracy of CNN-based model is higher than that of FCN + RF-based model. However, the two models have different characteristics of simulated urban perception scores, so it is necessary to analyze the difference and similarity between model scores and the specific application environment. Based on the app's street view data, we simulate the real environment and compare the two models. Figure 6 shows the distribution of six perceptions in the study area obtained from the two models.

From Fig. 6, we find that there is a strong similarity distribution between the two methods in the study area (except for beautiful scores). On the left side of the study area, the scores of wealthy, lively and safety are significantly higher than those of the right. The boring scores show a high level in the whole study area. When under the overpass of the city (the brown line area on the left in Fig. 6 e(1) and e(2)), the depressing scores increase significantly.

Urban streetscape can reflect spatial distribution of landscape elements in a certain area. The spatial distribution of the landscape is potentially related to regional

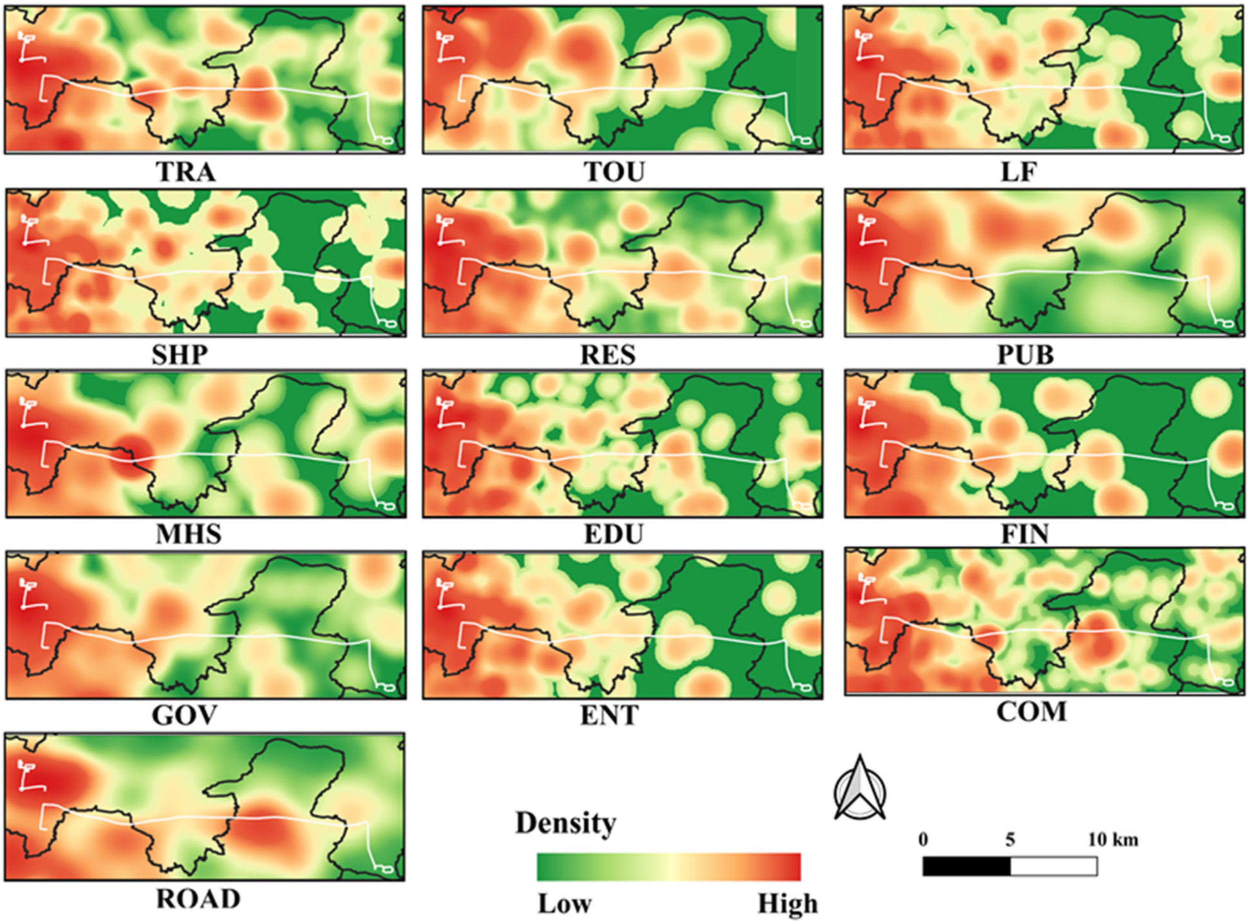

Fig. 5 Distribution of Gaode POIs and OSM in the study area using kernel density. POI categories: traffic facilities (TRA), tourist attractions (TOU), life service (LF), shopping (SHP), residential communities (RES), public facilities (PUB), medical service (MHS), education facilities (EDU), financial services (FIN), government (GOV), entertainment facilities (ENT), commercial and business (COM). ROAD indicates the density of the road network

land use and functional heterogeneity (Zhang et al., 2018a, b, c). Therefore, urban perception simulated by the street view is also affected by the land use and urban heterogeneity pattern to a certain extent. In order to further analyze the perception differences in real scenes, we selected several scenes with different land use and heterogeneity patterns (Fig. 7).

Urban central areas usually have a high scene complexity, which have diverse object types and mixed urban functions, and the area has a high heterogeneity pattern (Deng et al., 2020; Irwin & Bockstael, 2007). While urban suburb areas usually have a low scene complexity, where the objects are relatively homogeneous and the urban functions are simple (Zhou, Pickett, & Cadenasso, 2017). The suburb area has a low heterogeneity pattern (Irwin & Bockstael, 2007). The two models have evident differences in perception score between commercial center and suburb.

Table 1 Testing accuracy of the urban perception estimation via CNN

Perceptions	Pearson R	RMSE	MAE
Safety	0.921	5.207	2.941
Lively	0.930	7.656	4.942
Wealthy	0.928	5.968	3.536
Beautiful	0.939	6.441	3.462
Boring	0.900	5.590	2.814
Depression	0.926	5.194	2.647

Table 2 Testing accuracy of the urban perception estimation via FCN + RF

Perceptions	Pearson R	RMSE	MAE
Safety	0.787	8.273	3.954
Lively	0.797	11.163	5.883
Wealthy	0.798	9.537	4.810
Beautiful	0.839	9.950	6.302
Boring	0.820	7.270	4.528
Depression	0.793	8.411	4.543

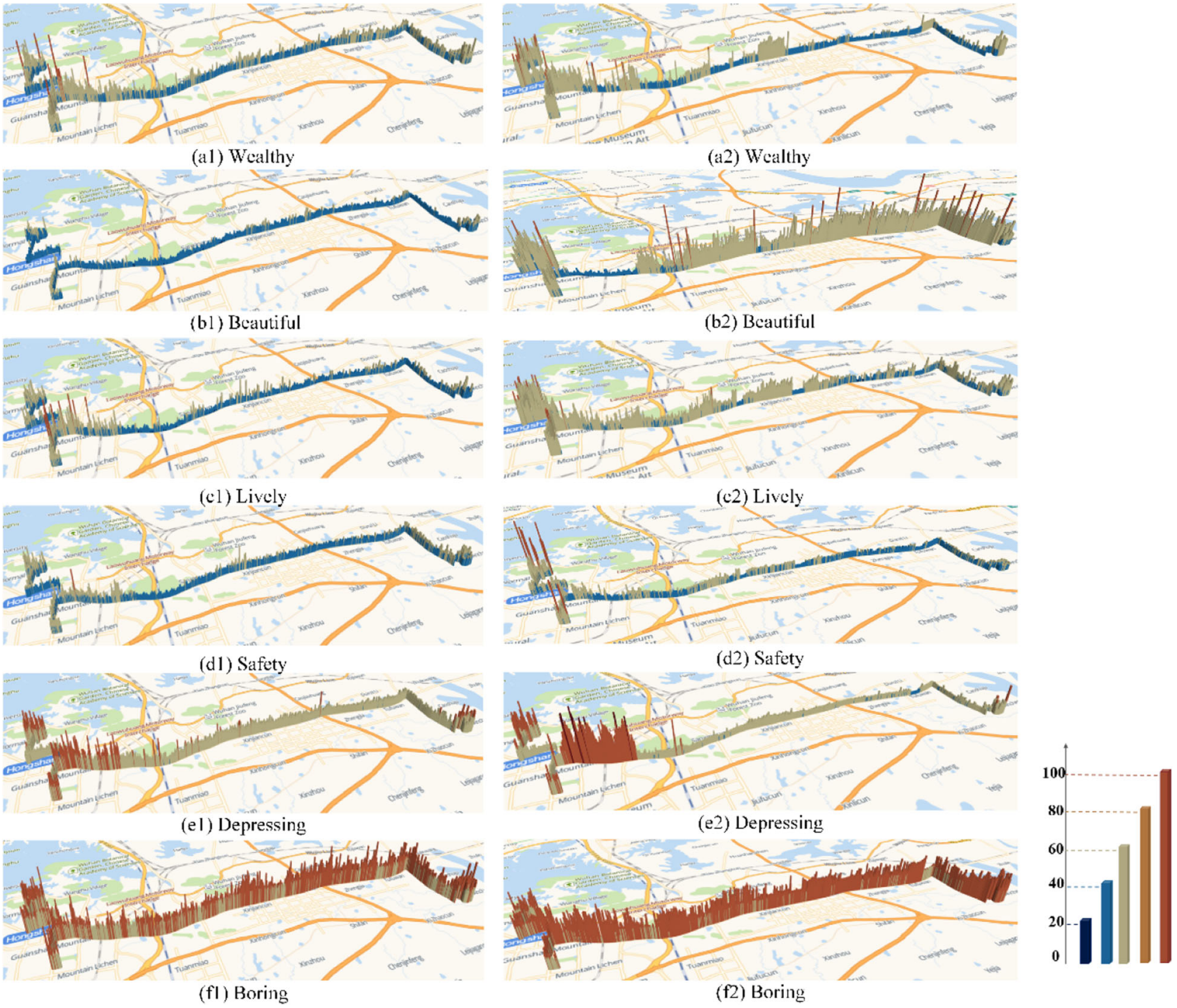

Fig. 6 The distribution of urban perception results in the study area. (a1) - (f1) is the perception result of CNN based model, and (a2) - (f2) is the perception result of FCN + RF based model. According to the different scores (the range of scores is 0–100), the results are divided into five equal parts. The cube height represents the perception score. [0,20] is dark blue, (20.40] is sky blue, (40,60] is yellow, (60.80] is orange, and (80,100] is brown

In the commercial center, the FCN + RF-based model is more consistent with the real scene in the perception of beautiful, lively and wealthy. However, the CNN-based model achieves more reasonable results in the suburb. The green plants and sky in the street scene are positive visual elements, which will have a positive impact on the feeling of beautiful, quiet and happy (Kaufman & Lohr, 2002; Quercia, O'Hare, & Cramer, 2014), but have a negative correlation with depressing (Helbich et al., 2019; Zhang et al., 2018a, b, c). (a1), (a3), (a4) and (a5) in Fig. 7 are mixed scene of complex landscape and natural landscape. Due to trees and other natural landscapes, people will have higher beautiful perception and lower depressing perception. In these scenes, FCN + RF-based model gets higher beautiful score. CNN-based model in Fig. 7 (a4) gives a significantly abnormal score for depressing perception. Contrary to (a3), (a4) and (a5) in Fig. 7, (b1) and (b4) in Fig. 7 are under an overpass on suburban roads, which blocks the sky and other elements, indicating that large-scale human-made features can have a negative impact (Zhang et al., 2018a, b, c). By extracting the image's spatial structure and color features, the CNN-based model gives a high depressing score, which is significantly higher than that of the FCN + RF-based model.

There is a close relationship between the traffic/crowd flow and the environment liveness (Yao et al., 2019). Downtown areas tend to have higher traffic and crowd flow (Zhang et al., 2018a, b, c; Zhang et al., 2019). For example, (a3), (a6), (a7) and (a9) in Fig. 7 are located on

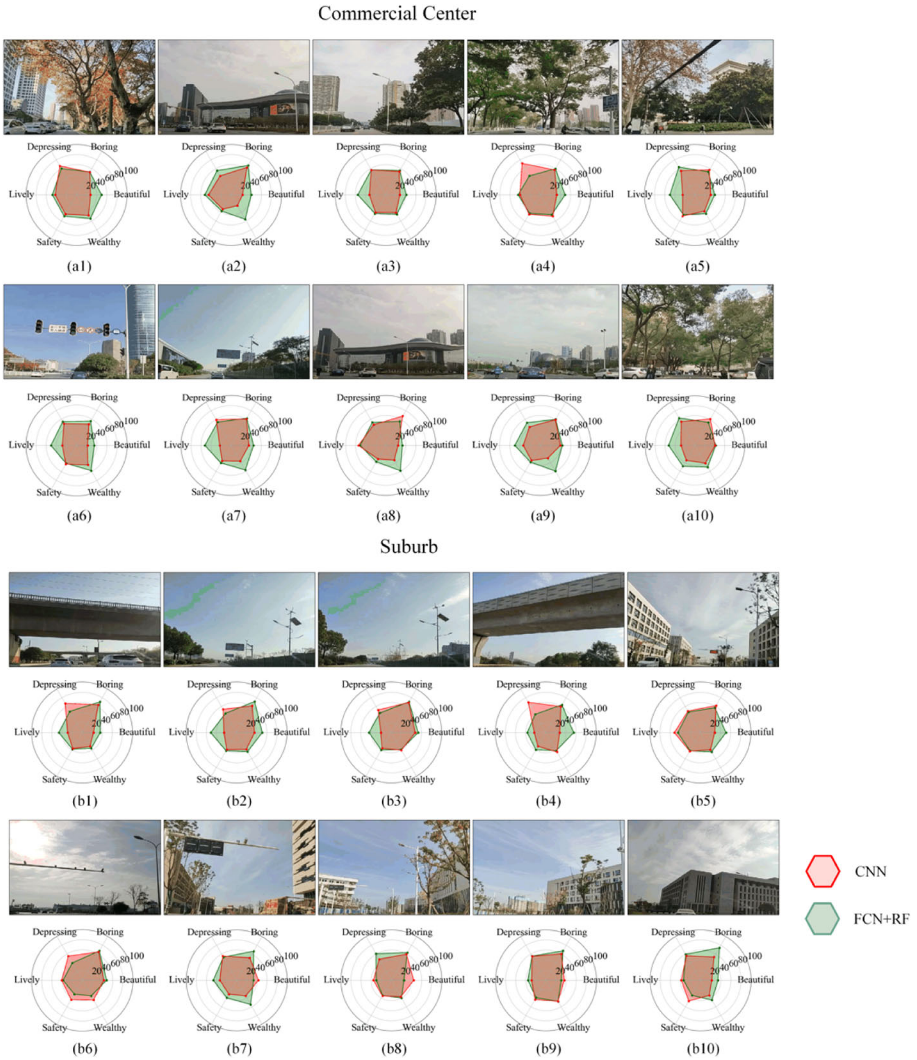

Fig. 7 The comparison of the perception scores by the CNN-based model and FCN + RF-based model for the samples in case study areas: (a1) -(a10) are scenes with strong spatial heterogeneity; (b1)- (b10) are scenes with weak spatial heterogeneity

the urban trunk road with dense traffic flow; (a5) and (a10) in Fig. 7 are located in the campus with dense pedestrian flow. The lively scores of the FCN + RF-based model in these scenes are higher than those of CNN-based model. However, in suburb, the objects are single, and the people and vehicles flow are rare, such as (b2) and (b3) in Fig. 7, which are scenes driving on a suburban highway, the FCN + RF-based model gives a high lively score, which is inconsistent with the real scene experience. The buildings density has a positive impact on the wealthy perception (Yao et al., 2019). The FCN + RF-based model gets higher wealthy scores in (a2) and (a8) which are complete landscape scenes in Fig. 7. This is because they are located in the city center's commercial area, and the prosperity of buildings is higher than that of other areas. The wealthy score of the FCN + RF-based model is more reasonable and effective.

4.3 Interpretability analysis of the model

There is a specific correlation between urban function and urban perception, and POI data can reflect urban functional areas (Hu & Han, 2019; Zhang, Du, & Wang, 2017). Therefore, based on POI and OSM, we analyze the model's interpretability fitting results from the perspective of urban function. Tables 3 and 4 are the urban perception results of the CNN-based model and FCN + RF-based model fitted by POI and OSM. The two models show good adaptability to each perception ($R^2 >$ 0.89, Pearson $R > 0.94$). Our results show that there is a strong nonlinear relationship between urban perception and urban function. POI and OSM can accurately estimate the distribution of urban perception and is an effective method to evaluate urban perception. By comparing the results of Tables 3 and 4, we find that the fitting accuracy of the CNN-based model is better than

Table 3 Testing accuracy of the urban perceptions (CNN) based on the POI and OSM densities via RF

Perceptions	R^2	Pearson R	RMSE
Wealthy	0.944	0.975	0.016
Beautiful	0.937	0.972	0.018
Lively	0.940	0.973	0.020
Safety	0.927	0.969	0.015
Depressing	0.940	0.973	0.014
Boring	0.928	0.969	0.015

that of the FCN + RF-based model (RMSE of CNN is 0.016, RMSE of FCN + RF is 0.021).

Yao et al. (2019) proves that FCN + RF-based model has good interpretability by using POI and feature weight ranking of random forest. This study will also refer to Yao et al. (2019) study to verify the interpretability of CNN-based model. The weight relationship between urban perception and POI or OSM categories is shown in Table 5.

Through the analysis of the results, we find that residential communities and roads are the most critical factors affecting urban perception in the case study area. In Yao et al. (2019), the six emotions are greatly affected by Edu and Gov. This study finds that the urban functions that affect urban perception are different from each other. ENT has a considerable weight of the wealthy perception, consistent with the general recognition that entertainment consumption places are in developed areas. Because depression is closely related to academic pressure, education areas have a great impact on depression perception (Ang & Huan, 2006). The perception of safety and lively have very strong relationships with the residential communities. This is because the community has the characteristics of frequent people flow and has reasonable security measures, which will bring great comfort to people (Holmberg, 2005).

5 Discussion

In this study, an end-to-end urban perception evaluation model based on street view images is proposed by building a multilayer CNN. We choose the FCN + RF-based model as the comparison method and analyze the difference in urban perception results between the two models in the study area. By combining POI data and OSM data, we calculate the weight of the driving factors that affect the urban perception, and analyze the interpretability of the results based on the CNN-based model.

Table 4 Testing accuracy of the urban perceptions (FCN + RF) based on the POI and OSM densities via RF

Perceptions	R^2	Pearson R	RMSE
Wealthy	0.942	0.971	0.020
Beautiful	0.946	0.973	0.020
Lively	0.899	0.948	0.026
Safety	0.945	0.972	0.026
Depressing	0.973	0.986	0.016
Boring	0.893	0.945	0.022

Both models achieved ideal prediction accuracy (Pearson $R > 0.78$). However, the CNN-based model has better accuracy than the FCN + RF based model. The RMSE index of the CNN-based model is 2.6 lower than that of the FCN + RF-based model. Therefore, the result of the CNN-based model is slightly better than that of the FCN + RF-based model. In addition, compared with the method of using semantic segmentation first and then using random forest to determine the perception scores, the CNN-based model directly obtains the perception scores by inputting the images, which is faster and easier.

By extracting the high-dimensional features, the CNN-based model can obtain a high degree of nonlinear correlation between urban perception and urban functions. Through the study of the correlation and the weight between urban function and urban perception, we find that there is spatial similarity between the distribution of urban perception and the distribution of urban functional areas. POI data and OSM data can accurately estimate the distribution of urban perception (the average error of RMSE for six perceptions is 0.016). The case study quantificationally determines the impact of the urban functions on urban perception and proves that the CNN-based model proposed in this study has better rationality in promoting the evaluation of local urban perception. The FCN + RF-based model ignores the spatial topological features of the ground objects, which leads to its lower correlation with urban function than the CNN-based model (the average error of RMSE of six perceptions is 0.021).

The CNN-based model is more suitable for scenes with weak spatial heterogeneity, such as small cities or suburbs in central China. The FCN + RF-based model has more advantages for urban areas with strong spatial heterogeneity, such as the developed metropolises in central China. The CNN-based model can extract the features of the ground objects, such as color, texture and density (Hu et al., 2015; Jiao et al., 2017; Kim & Pavlovic, 2016; Sahiner et al., 1996). These detailed features are more reasonable in representing scenes with weak spatial heterogeneity. The FCN + RF-based model is suitable for scenes with strong spatial heterogeneity. When scenes contain strong spatial heterogeneity, the features of the ground objects become fuzzy, and the impact on urban perception will decrease. By directly calculating the ratio of the ground objects, a more accurate score can be obtained. Therefore, the two models have their

Table 5 Fitting weights perceptions and POI or OSM categories: government (GOV), life service (LF), medical service (ME), public facilities (PUB), residential communities (RES), traffic facilities (TRA), financial services (FIN), entertainment facilities (ENT), tourist attractions (TOU), road (RO), commercial and business (COM), shopping (SHP), and education facilities (EDU). A gradient background color from blue to yellow to red indicates a gradual increase in value

Category	Wealthy	Beautiful	Lively	Safety	Depressing	Boring
GOV	0.039	0.045	0.048	0.056	0.132	0.078
LF	0.061	0.032	0.035	0.058	0.039	0.044
ME	0.042	0.070	0.057	0.057	0.042	0.079
PUB	0.042	0.067	0.054	0.079	0.065	0.075
RES	0.125	0.081	0.149	0.154	0.142	0.077
TRA	0.058	0.072	0.079	0.063	0.066	0.115
FIN	0.071	0.063	0.067	0.042	0.052	0.054
ENT	0.120	0.032	0.059	0.052	0.063	0.052
TOU	0.156	0.078	0.063	0.066	0.062	0.155
RO	0.075	0.251	0.192	0.174	0.062	0.076
COM	0.063	0.070	0.058	0.071	0.069	0.063
SHP	0.054	0.058	0.049	0.045	0.053	0.055
EDU	0.093	0.080	0.091	0.083	0.153	0.076

own advantages. The method needs to be chosen according to the actual spatial heterogeneity of the urban environment to obtain a more accurate urban perception score.

There are still some deficiencies and many opportunities for future study. First, the mobile video is shot along the city streets. Considering that the front view plays a leading role in commuting, this study only selects the front view of the car and does not take a left or right view into account. However, the areas beside the road, such as parks and communities, also have a certain impact on people's cognition (Abkar et al., 2010; Oetzel et al., 2011). The purpose of this study is to analyze the applicability of the two models. Forward-looking images can effectively explain the rationality of the results of urban perception. Therefore, in future studies, we will consider adding two side views and more street views on different blocks; expanding our study to actual application scenarios will also be considered.

Second, urban perception is related not only to streetview but also to other factors in the city, such as season, temperature, humidity and noise (Gunnarsson et al., 2017; Hong & Jeon, 2015). Therefore, in future work, more evaluation objectives will be considered in the study of urban perception, and an end-to-end perception model will be adopted to obtain higher accuracy and stronger interpretability.

The third is what we need to do in the future. At present, there is no Chinese urban perception data set (official). Therefore, when comparing different methods, we can only compare and analyze the existing small-scale urban data. In the future, we will collect street view images of other regions (urban, suburban, rural) based on the collection method proposed in this study and construct Chinese perception data set to increase the persuasiveness.

6 Conclusion

In view of the consistency and interpretability of the prediction results of the two different urban perception models, this study extracts the street view images in the study area through mobile video, constructs a CNN-based model and FCN + RF-based model by using the China urban perception dataset, and compares the results of the different models on two urban road networks in Wuhan, China. In this study, the prediction accuracy is ideal (the Pearson R of the CNN-based model is 12% higher than that of the FCN + RF-based model, and the RMSE of the CNN-based model is 2.6% lower than that of the FCN + RF-based model). This shows that the proposed CNN-based model is effective for urban perception assessment. The two models both have reached the ideal accuracy. By using POI data and OSM data for auxiliary analysis, we find that there is a strong nonlinear correlation between urban function and urban perception. The CNN-based model has more advantages in predicting the urban perception by extracting the high-dimensional features and has a higher degree of correlation with the urban function. The scene suitability of the two models is different. The

CNN-based model is suitable for scenes with weak spatial heterogeneity, and the FCN + RF-based model is suitable for scenes with strong spatial heterogeneity. This study can accurately and quickly identify the exposed perception of residents and will promote urban planners to integrate the concept of urban perception into the planning practice. The results will provide decision support for government managers in urban planning to achieve a more sustainable and human-oriented urban development.

Authors' contributions

Ruifan Wang and Shuliang Ren completed the experiment and wrote the paper. Jiaqi Zhang revised the paper. Yao Yao provided the idea and funding, and revised the paper. Yu Wang completed the comparison experiment and model. Qingfeng Guan revised the paper. The authors read and approved the final manuscript.

Funding

This work was supported by the National Key R&D Program of China (Grant No. 2019YFB2102903), the National Natural Science Foundation of China (Grant No. 41801306 and 41671408) and the Natural Science Foundation of Hubei Province (Grant No. 2017CFA041).

Availability of data and materials

All of the data and code involved in this study can be downloaded at https://doi.org/10.6084/m9.figshare.12552233.

Competing interests

No conflict of interest exists.

Received: 27 November 2020 Accepted: 1 February 2021
Published online: 29 March 2021

References

Abkar, M., et al. (2010). The role of urban green spaces in mood change. *Australian Journal of Basic and Applied Sciences, 4*(10), 5352–5361.

Ang, R. P., & Huan, V. S. (2006). Relationship between academic stress and suicidal ideation: Testing for depression as a mediator using multiple regression. *Child Psychiatry and Human Development, 37*(2), 133.

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

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(12), 2481–2495.

Been, V., et al. (2016). Preserving history or restricting development? The heterogeneous effects of historic districts on local housing markets in New York City. *Journal of Urban Economics, 92*, 16–30.

Blaxland, M., Shang, X., & Fisher, K. R. (2014). Introduction: People oriented: A new stage of social welfare development in China. *Journal of Social Service Research, 40*(4), 508–519.

Bulat, A., & Tzimiropoulos, G. (2016). Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision* (pp. 717–732). Springer, Cham.

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

Cordts, M., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3213–3223).

Deng, L., et al. (2017). CNN based semantic segmentation for urban traffic scenes using fisheye camera. In *IEEE Intelligent Vehicles Symposium (IV)* (pp. 231–236).

Deng, Y., et al. (2020). Geographical transformations of urban sprawl: Exploring the spatial heterogeneity across cities in China 1992–2015. *Cities, 105*, 102415.

Dubey, A., et al. (2016). Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision* (pp. 196–212). Springer, Cham.

Gunnarsson, B., et al. (2017). Effects of biodiversity and environment-related attitude on perception of urban green space. *Urban Ecosystem, 20*(1), 37–49.

Ha, I., et al. (2018). Image retrieval using BIM and features from pretrained VGG network for indoor localization. *Building and Environment, 140*(2018), 23–31.

Han, Z., et al. (2020). Comparing fully deep convolutional neural networks for land cover classification with high-spatial-resolution Gaofen-2 images. *ISPRS International Journal of Geo-Information, 9*(8), 478.

Harvey, C., et al. (2015). Effects of skeletal streetscape design on perceived safety. *Landscape and Urban Planning, 142*, 18–28.

He, K., et al. (2016). Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Helbich, M., et al. (2019). Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environment International, 126*, 107–117.

Holmberg, L. (2005). Policing and the feeling of safety: The rise (and fall?) of community policing in the Nordic countries. *Journal of Scandinavian Studies in Criminology and Crime Prevention, 5*(2), 205–219.

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.

Hu, C., et al. (2015). Vehicle color recognition with spatial pyramid deep learning. *IEEE Transactions on Intelligent Transportation Systems, 16*(5), 2925–2934.

Hu, Y., & Han, Y. (2019). Identification of urban functional areas based on POI data: A case study of the Guangzhou economic and technological development zone. *Sustainability, 11*(5), 1385.

Ioffe, S. and Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Irwin, E. G., & Bockstael, N. E. (2007). The evolution of urban sprawl: Evidence of spatial heterogeneity and increasing land fragmentation. *Proceedings of the National Academy of Sciences, 104*(52), 20672–20677.

Jiao, L., et al. (2017). Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*(10), 5585–5599.

Kaufman, A. J., & Lohr, V. I. (2002). Does plant color affect emotional and physiological responses to landscapes? In *XXVI International Horticultural Congress: Expanding Roles for Horticulture in Improving Human Well-Being and Life Quality* (pp. 229–233).

Kim, J., & Pavlovic, V. (2016). A shape-based approach for salient object detection using deep learning. In *European Conference on Computer Vision* (pp. 455–470). Springer, Cham.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.

Liu, X., et al. (2017a). Place-centric visual urban perception with deep multi-instance regression. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 19–27).

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

Liu, X., et al. (2018). Classifying high resolution remote sensing images by fine-tuned VGG deep networks. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 7137–7140).

Liu, Y., et al. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*(3), 512–530.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–343').

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*(2), e0171110.

Lu, X., et al. (2017). Feature extraction and fusion using deep convolutional neural networks for face detection. *Mathematical Problems in Engineering, 2017*, 2017.

Middel, A., et al. (2019). Urban form and composition of street canyons: A human-centric big data and deep learning approach. *Landscape and Urban Planning, 183*, 122–132.

Monti, R. P., Tootoonian, S., & Cao, R. (2018). Avoiding degradation in deep feed-forward networks by phasing out skip-connections. In *International Conference on Artificial Neural Networks* (pp. 447–456). Springer, Cham.

Naik, N., Raskar, R., & Hidalgo, C. E. S. A. (2016). Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *American Economic Review, 106*(5), 128–132.

Naik, N., et al. (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).

Oetzel, J., et al. (2011). Creating an instrument to measure people's perception of community capacity in American Indian communities. *Health Education & Behavior, 38*(3), 301–310.

Ordonez, V., & Berg, T. L. (2014). Learning high-level judgments of urban perception. In *European Conference on Computer Vision* (pp. 494–510). Springer, Cham.

Ozkan, U. Y. (2014). Assessment of visual landscape quality using IKONOS imagery. *Environmental Monitoring and Assessment, 186*(7), 4067–4080.

Palczewska, A., et al. (2014). Interpreting random forest classification models using a feature contribution method. In *Integration of Reusable Systems* (pp. 193–218). Springer, Cham.

Piramanayagam, S., et al. (2016). *Classification of remote sensed images using random forests and deep learning framework*. International Society for Optics and Photonics, *10004,* 100040L.

Porzi, L., et al. (2015). Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 139–148).

Quercia, D., O'Hare, N. K., & Cramer, H. (2014). Aesthetic capital: What makes London look beautiful, quiet, and happy? In *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing* (pp. 945–955).

Sahiner, B., et al. (1996). Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. *IEEE Transactions on Medical Imaging, 15*(5), 598–610.

Salesses, P., Schechtner, K., & Hidalgo, C. E. S. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PLoS One, 8*(7), e68400.

Sampson, R. J. (2012). *Great American city: Chicago and the enduring neighborhood effect*. University of Chicago Press.

Shin, H., et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging, 35*(5), 1285–1298.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

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

Sun, Z., Li, F., & Huang, H. (2017). Large scale image classification based on CNN and parallel SVM. In *International conference on neural information processing* (pp. 545–555). Springer, Cham.

Wang, R., et al. (2019a). Using street view data and machine learning to assess how perception of neighborhood safety influences urban residents' mental health. *Health & Place, 59*, 102186.

Wang, R., et al. (2019b). 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.

Weber, R., Schnier, J. O. R., & Jacobsen, T. (2008). Aesthetics of streetscapes: Influence of fundamental properties on aesthetic judgments of urban space. *Perceptual and Motor Skills, 106*(1), 128–146.

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.

Yao, Y., et al. (2019). A human-machine adversarial scoring framework for urban perception assessment using street-view images. *International Journal of Geographical Information Science, 33*(12), 2363–2384.

Yuan, J., Zheng, Y., & Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 186–194).

Zamani Joharestani, M., et al. (2019). PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere, 10*(7), 373.

Zhai, Y., et al. (2020). Simulating urban land use change by integrating a convolutional neural network with vector-based cellular automata. *International Journal of Geographical Information Science, 34*(7), 1475–1499.

Zhang, F., et al. (2018a). Representing place locales using scene elements. *Computers, Environment and Urban Systems, 71*, 153–164.

Zhang, F., et al. (2018b). Framework for virtual cognitive experiment in virtual geographic environments. *ISPRS International Journal of Geo-Information, 7*(1), 36.

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

Zhang, F., et al. (2019). Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS Journal of Photogrammetry and Remote Sensing, 153*, 48–58.

Zhang, X., Du, S., & Wang, Q. (2017). Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. *ISPRS Journal of Photogrammetry and Remote Sensing, 132*, 170–184.

Zhou, B., et al. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 633–641).

Zhou, B., et al. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision, 127*(3), 302–321.

Zhou, H., et al. (2016). Image semantic segmentation based on FCN-CRF model. In *2016 International Conference on Image, Vision and Computing (ICIVC)* (pp. 9–14). IEEE, Portsmouth.

Zhou, W., Pickett, S. T. A., & Cadenasso, M. L. (2017). Shifting concepts of urban spatial heterogeneity and their implications for sustainability. *Landscape Ecology, 32*(1), 15–30.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.