Landscape and Urban Planning 180 (2018) 148–160

Contents lists available at ScienceDirect

Landscape and Urban Planning

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

Research Paper

Measuring human perceptions of a large-scale urban region using machine learning

Fan Zhang[a,b,c], Bolei Zhou[d], Liu Liu[e], Yu Liu[a], Helene H. Fung[f], Hui Lin[b,g,*], Carlo Ratti[c]

[a] Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing 100871, China
[b] Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong, China
[c] Senseable City Laboratory, Massachusetts Institute of Technology, MA 02139, USA
[d] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, MA 02139, USA
[e] China Academy of Urban Planning and Design Shanghai Branch, Shanghai 200335, China
[f] Department of Psychology, The Chinese University of Hong Kong, Hong Kong, China
[g] College of Geography and Environment, Jiangxi Normal University, Nanchang 330200, China

ARTICLE INFO

Keywords:
Urban perception
Place semantics
Street-level imagery
Deep learning
Built environment

ABSTRACT

Measuring the human sense of place and quantifying the connections among the visual features of the built environment that impact the human sense of place have long been of interest to a wide variety of fields. Previous studies have relied on low-throughput surveys and limited data sources, which have difficulty in measuring the human perception of a large-scale urban region at flexible spatial resolutions. In this work, a data-driven machine learning approach is proposed to measure how people perceive a place in a large-scale urban region. Specifically, a deep learning model, which has been trained on millions of human ratings of street-level imagery, was used to predict human perceptions of a street view image. The model achieved a high accuracy rate in predicting six human perceptual indicators, namely, safe, lively, beautiful, wealthy, depressing, and boring. This model can help to map the distribution of the city-wide human perception for a new urban region. Furthermore, a series of statistical analyses was conducted to determine the visual elements that may cause a place to be perceived as different perceptions. From the 150 object categories segmented from the street view images, various objects were identified as being positively or negatively correlated with each of the six perceptual indicators. The results take researchers and urban planners one step toward understanding the interactions of the place sentiments and semantics.

1. Introduction

For decades, as human settlement in modern cities has changed and been reshuffled, places in the cities have accordingly been reshaped and gradually grown unequally in terms of their location, physical setting, and the groups of people that live in them (Salesses, Schechtner, & Hidalgo, 2013). Place, defined as "…spatial locations that have been given meaning by human experiences" (Tuan, 1977), has been a fundamental component of everyday life (Morison, 2002) and has influenced people's cognition and perceptions in the stream of experience (Goodchild, 2011; Tuan, 1977). In particular, the sense of place refers to human perceptions and nebulous meanings associated with a place. Measuring the human sense of place can potentially enrich place semantics, which will further help researchers understand the underlying urban heterogeneity patterns and reveal the impacts of urban function. Indeed, learning how to gather knowledge about physical settings and the visual information of a place that affects the experience of observers has long been of interest to a wide variety of fields (Kaplan & Kaplan, 1989; Lynch, 1960; Nasar, 1997; Tuan, 1977). Previous studies were conducted by the traditional data collection methods, such as interviews and questionnaires (Cresswell, 1992; Montello, Goodchild, Gottsegen, & Fohl, 2003), which are laborious, costly and time-consuming. These methods constrain the scalability to a small research area. The size of modern cities increases and the physical appearance of cities changes rapidly. Developing a comprehensive body of knowledge about the streets, places, and cities is becoming even more difficult but more meaningful than ever before. It is of great significance for researchers, urban planners and decision makers to understand how citizens perceive and evaluate places in a large-scale urban region at a high resolution. This work makes a contribution by enhancing the understanding of human perceptions of places in a large-scale urban environment in an automatic and efficient way by using machine learning

* Corresponding author at: Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong, China.
E-mail address: huilin@cuhk.edu.hk (H. Lin).

https://doi.org/10.1016/j.landurbplan.2018.08.020
Received 20 October 2017; Received in revised form 20 August 2018; Accepted 24 August 2018
Available online 13 September 2018

and street-level imagery.

Due to the rapid development of map services and volunteered geographic information (VGI) (Anguelov et al., 2010; Goodchild, 2007; Liu et al., 2015), a massive amount of geo-tagged images have been compiled and made publicly available that can describe every corner of a city. Street-level imagery has enabled the development of new approaches to observe, perceive, and understand the built environment. In addition, great progress has been made on recent advance of computer vision technique for recognizing the image content by deep learning, which has attracted much attention and achieved great success in multiple fields due to its powerful ability in automatic image feature learning and representation (Hinton et al., 2012; He, Gkioxari, Dollár, & Girshick, 2017; LeCun, Bengio, & Hinton, 2015).

In this study, a deep learning based approach is first proposed to model and predict human perceptions of the physical setting of a place. The approach is able to predict the six human perceptual indicators accurately, namely, safe, lively, boring, wealthy, depressing, and beautiful for a new urban region. Second, we investigate the connections between urban visual elements and perceptions using multivariate regression analyses, and tried to determine "which visual elements may cause a place to be perceived as a specific perception". Various visual elements were identified to be positively or negatively correlated with perceptual indicators through statistical analyses. The results take researchers and urban planners one step toward understanding place sentiments and semantics by exploring underlying urban heterogeneity patterns and revealing the impacts of urban function.

The remainder of this paper is organized as follows. In Section 2, we review the related work. In Section 3, we introduce the two massive geo-tagged image datasets used in this work, the MIT Place Pulse dataset and the street view images collected from two cities in China. Section 4 introduces the approach used to predict human perceptions of street view images and the approach used to determine the visual elements that are related to human perceptions. Section 5 describes the experiment and the results of the proposed method. In Sections 6 and 7, we discuss the significance and limitations of this study and draw some conclusions.

2. Related work

2.1. Measuring human perceptions

The sense of place refers to human perceptions and nebulous meanings based on our prior experience with a place (Cresswell, 2014; Tuan, 1977). For decades, a wide variety of disciplines and fields including geography, urban planning, environmental psychology and neuroscience have considered about the connections between the environment and human perceptions (Kaplan & Kaplan, 1989; Lynch, 1960; Nasar, 1997; Tuan, 1977).

The seminal work of Tuan, *Space and Place: The Perspective of Experience* (Tuan, 1977) focused on how space and place are formed and how the feelings about place are affected. Regarding urban planning, the literature has focused mostly on the built environment of cities. Lynch identified three components that constitute an individual's feelings about the environment: identity, structure, and meaning, of which meaning indicates the practical and perceptual value of the place to the individual (Lynch, 1960). Rachel and Stephen Kaplan's work paid more attention to understanding the effect of nature on people's perceptions and mental health from the perspective of environmental psychology (Kaplan & Kaplan, 1989). Similarly, Ulrich's research demonstrated that the natural environment is able to induce people's aesthetic and affective responses (Ulrich, 1983), which would have a restorative influence on patients (Ulrich, 1984).

Since these works were published, measuring the sense of place has become a research area that has been receiving increased attention. Studies have been conducted using surveys including interviews and questionnaires to measure certain evaluative dimensions, for instance, inviting subjects to rate the physical setting of a place using a 1–10 scale (Michael, 2005; Nasar, 1997; Schroeder & Anderson, 1984). Due to the development of neuroscience, electroencephalographs (EEGs) and functional Magnetic Resonance Imaging (fMRI) have been employed to measure human emotions by the response of brain signals to different visual setting of the environment (Valtchanov & Ellard, 2015; Mallgrave, 2010; Mišić et al., 2014). Quercia et al. conducted an on-line survey for 700,000 streets to collect data on the sense of street in terms of safety and beauty (Quercia, O'Hare, & Cramer, 2014).

In recent years, the proliferation of crowdsourcing technology has enhanced the ability to collect a massive amount of images to represent the physical setting of place and to predict human perceptual responses of images. Research efforts have been made to predict the memorability (Isola, Xiao, Torralba, & Oliva, 2011), virality (Deza & Parikh, 2015), city/object style (Doersch, Singh, Gupta, Sivic, & Efros, 2012; Jae Lee, Efros, & Hebert, 2013), aesthetics and interestingness of street scenes (Datta, Joshi, Li, & Wang, 2006; Dhar, Ordonez, & Berg, 2011; Machajdik & Hanbury, 2010), among others.

2.2. Representing the physical setting of a place using street-level imagery

Over the years, the rapid development of map services (Anguelov et al., 2010) and volunteered geographic information (VGI) (Goodchild, 2007) has provided a massive amount of geo-tagged images. This new source of data has can provide information on every corner of a city and has been enabling broader and more in-depth quantitative research in related fields. These data enhance the understanding of the city's physical and dynamic characteristics by detecting landmark (Hays & Efros, 2015), recognizing urban identities (Liu, Zhou, Zhao, & Ryan, 2016; Zhang, Zhang, Liu, & Lin, 2018), evaluating the inequality of the living environment (Salesses et al., 2013), and modeling human activities (Arase, Xie, Hara, & Nishio, 2010) and popular places (Crandall, Backstrom, Huttenlocher, & Kleinberg, 2009). These new data also provide information on the physical and social structures of dynamic urban environments (Crandall et al., 2009; Less et al., 2015).

In 2013, the MIT Media Lab initiated the program "Place Pulse", which is a data collection platform that enables volunteers to participate in the urban perception rating experiment. By the end of 2016, the MIT Place Pulse dataset had collected 1,170,000 pairwise comparisons from 81,630 online participants for 110,988 cityscape images. Inspired by the dataset and enabled by the recent progress in machine learning techniques, a number of studies have been conducted to analyze human perceptions of urban appearance (Dubey, Naik, Parikh, Raskar, & Hidalgo, 2016; Glaeser, Kominers, Luca, & Naik, 2016; Naik, Philipoom, Raskar, & Hidalgo, 2014Ordonez and Berg, 2014; Salesses et al., 2013).

However, previous approaches have difficulty in extracting high-level information about the natural image because they use low/mid-level image features including Gist, SIFT- Fisher Vectors, DeCAF features (Ordonez & Berg, 2014), geometric classification map, color Histograms, HOG2x2, and Dense SIFT. Naik et al. (2014). With regard to building models to predict image labels, Support Vector Machine (SVM) and Linear Regression (LR) were used in Ordonez and Berg (2014), Support Vector Regression was used in Naik et al. (2014), RankingSVM was used in Porzi, Rota Bulò, Lepri, and Ricci (2015), and several convolutional neural network based approaches were used in Ordonez and Berg (2014), Porzi et al. (2015) and Dubey et al. (2016). Among the various image representations and models, Deep Convolutional Neural Network (DCNN)- based approaches have outperformed conventional methods to a large extent (Dubey et al., 2016). This study introduces a DCNN model that is based on the Deep Residual Network (ResNet) (He, Zhang, Ren, & Sun, 2016), which won 1st place in the ImageNet Large Scale Visual Recognition Competition (Russakovsky et al., 2015).

2.3. Extracting high-level information from natural images using deep learning

Recently, due to the rapid development of high-performance computing systems and the availability of large-scale annotated datasets, a hierarchical and shift-invariant model has emerged in the form of DCNN. Due to its powerful ability in automatic image feature learning and representation, this model has attracted much attention and achieved great success in multiple fields, including speech recognition (Hinton et al., 2012), natural language processing (Sutskever, Vinyals, & Le, 2014), and visual object detection (Ren, He, Girshick, & Sun, 2015; He et al., 2017). In this study, DCNN was employed to conduct human perception modeling and prediction.

A very deep convolutional neural network is difficult to train and optimize because of vanishing gradients and the curse of dimensionality (Glorot & Bengio, 2010; He et al., 2016). ResNet is believed to be a good attempt to address this problem. ResNet was designed to learn the residual functions with regard to the layer inputs rather than learning the unreferenced functions (He et al., 2016). More specifically, we use the ResNet50 that was pre-trained on Places2 (http://places2.csail.mit.edu/) which is an image database that contains 10 million well-labeled images (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017). We also extracted the high-dimensional deep feature of each street view image from the Place Pulse database.

3. Massive geo-tagged image datasets

Two data sources were used in this study, of which (1), the MIT Place Pulse dataset was used to train the DCNN model to predict human perception, and (2), the street view images were used for predicting human perception in new urban regions.

3.1. MIT places pulse dataset

In 2013, the MIT Media Lab launched the project “Place Pulse 2.0”, an online data collection platform for collecting human perceptual ratings of urban appearance. On the website, the participants are shown two street view images that are randomly sampled from one city side-by-side; then, they are asked to choose one of the images as the response to questions such as the following:“which place looks more X?”, where the X can be one of six dimensions:“Safe”, “beautiful”, “depressing”, “lively”, “wealthy”, and “boring”. The participants select one response from three options, the left image, the right image or “equal”, to indicate their perceptual judgment. Fig. 1 shows the user interface of this platform.

The image dataset contains 110,988 street view images captured between 2007 and 2012, spanning 56 cities in 28 countries across 6 continents (as Table 1 lists). Fig. 2 depicts the geographic distribution of the image data in the MIT Place Pulse dataset. In terms of the scale of cities, ecumenopolis such as New York and London, as well as cities such as Glasgow and Gaborone, have been included in the dataset. For each city, the locations were densely and randomly sampled from the spatial region of the city. The meta-data of these images are also included in the dataset, including the geo-coordinates and camera heading degree. By October 2016, 1,169,078 pairwise comparisons had been collected from 81,630 online participants (Dubey et al., 2016).

Evaluation of the diversity, consistency and potential biases of the dataset. Volunteers from 162 developed and developing countries participated in this experiment, which indicates the diversity of the dataset. To explore the potential biases in the collection that may come from the demographics of the volunteers, a correlation significance test was conducted. The results indicated that there were no significant biases for groups with different demographics (Dubey et al., 2016; Salesses et al., 2013). Furthermore, the internal consistency of the ratings in the dataset was also tested by looking at the inter-user reproducibility and transitivity; both were found to be high (Salesses et al., 2013).

Perception score calculation. The rating data in the Place Pulse dataset were in the form of two-image comparisons. Even though a previous attempt was made to design a model that can be trained with comparison pairs (Dubey et al., 2016), we believe that using labeled single-samples is more practical in application, because in a real case we always want to know the observers' perceptions of one single scene rather than a comparison of two scenes. In this case, each image sample i might have been compared with other images i' several times, and intuitively, the percentage of times that i is selected essentially indicates the intensity of the perception for the specific dimension. In addition, the intensity of i' should also be considered and weighted when calculating the intensity of I.

Attempts have been made to parameterize this process based on intuition by Salesses et al. (2013) and Ordonez and Berg (2014) using strength of schedule methods. Similarly, Dubey et al. (2016) adopted the Microsoft Trueskill algorithm (Herbrich, Minka, & Graepel, 2007) to achieve this. In this case, we employed the former approach. First, we defined the positive rate (P) and the negative rate (N) of image i along a particular perceptual indicator as:

$$P_i = \frac{p_i}{p_i + e_i + n_i} \tag{1}$$

$$N_i = \frac{n_i}{p_i + e_i + n_i} \tag{2}$$

where p_i and n_i denote the number of times that image i was selected or not selected in the comparisons, respectively, and e_i refers to the number of times that image i was believed to be equal to another image in the comparison. Consequently, we were able to define the Q-score for image i along the specific perceptual indicator as:

$$Q_i = \frac{10}{3}\left(P_i + \frac{1}{p_i}\sum_{k_1=1}^{p_i} P_{k_1} - \frac{1}{n_i}\sum_{k_2=1}^{n_{k_2}} N_{k_2} + 1\right) \tag{3}$$

where k_1 and k_2 indicate the number of times that image i was selected and not selected, respectively, in the comparisons. According to Eq. 3, the final Q-score is actually the positive rate P_i corrected by the $P_{i'}$ and $N_{i'}$ of the images that it was compared with. Referring to previous studies in visual assessment (Nasar, 1997; Salesses et al., 2013), the score will be scaled to a range from 0 to 10 by adding the constant value 1 and multiplying the equation by $\frac{10}{3}$ to the equation.Fig. 3 shows 4 image samples with their corresponding perceptual scores (Q score) for the six dimensions. We can see that different types of urban scenes may induce different human perceptions. For example, the first image looks boring but lively, because we can identity the human activities in the scene. This image is also believed to be unsafe and not beautiful.

3.2. Street view images of Beijing and Shanghai

To predict the human perception of a new urban region, we also collected street view images from new cities. In this case, the image data of Shanghai and Beijing were obtained from Tencent Street View service (https://map.qq.com/) through an API. At an interval of 50 meters, the sampling locations were generated along the road network. The locations were used to request street view images. For each location point, the detailed request parameters were set as follows: image size: 480x600, compass heading of the camera: 0, 90, 180, 270 degrees; and the horizontal field of view of the image: 90 degrees.

In total, 245,388 images were collected from Shanghai and 135,175 images were collected from Beijing.

4. Deep learning of street view images to assess urban design

4.1. Modeling human perception

Modeling human perception requires training a machine learning

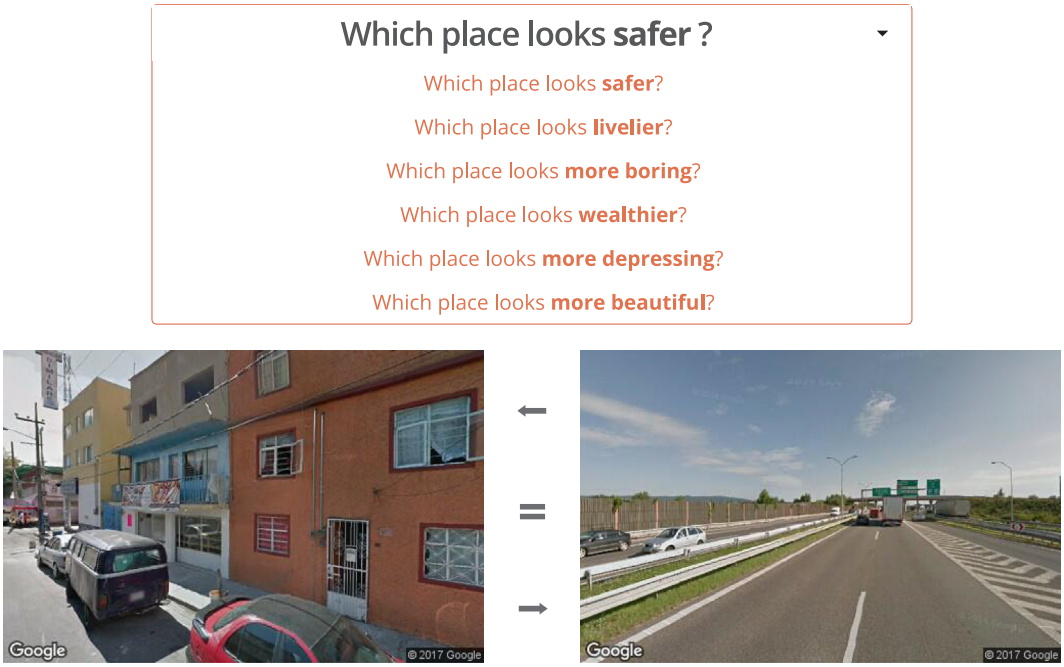

Fig. 1. The user interface of the MIT Place Pulse data collection platform (http://pulse.media.mit.edu/). Participants are asked to choose one of the two images in response to one of the six questions. Millions of human perception responses for the images have been collected.

Table 1
Statistics of image data in the MIT Place Pulse dataset. Images span 56 cities across 6 continents (the stats collected here were obtained from Dubey et al. (2016)).

Continent	#Cities	#Images
Asia	7	11,342
Africa	3	5,069
Australia	2	6,082
Europe	22	38,636
North America	15	33,691
South America	7	16,168
Total	56	110,988

model to predict the human perception score of one street view image along six perceptual indicators, namely, safe, lively, boring, wealthy, depressing, and beautiful. Fig. 4 illustrates the overview of human perception modeling and prediction. We formulate the human perception prediction problem as a binary classification task. In other words, we need to predict whether one image would be perceived as positive or negative. Handling the perceptual task with a binary classification model rather than a regression model is more practical, such as those used to predict aesthetics (Datta et al., 2006; Datta, Li, & Wang, 2008; Dhar et al., 2011), and urban perceptions (Ordonez & Berg, 2014). This method is more practical because human perceptual evaluation is highly uncertain and unstable, especially around the middle scores. The models used to predict the six perceptual indicators were trained

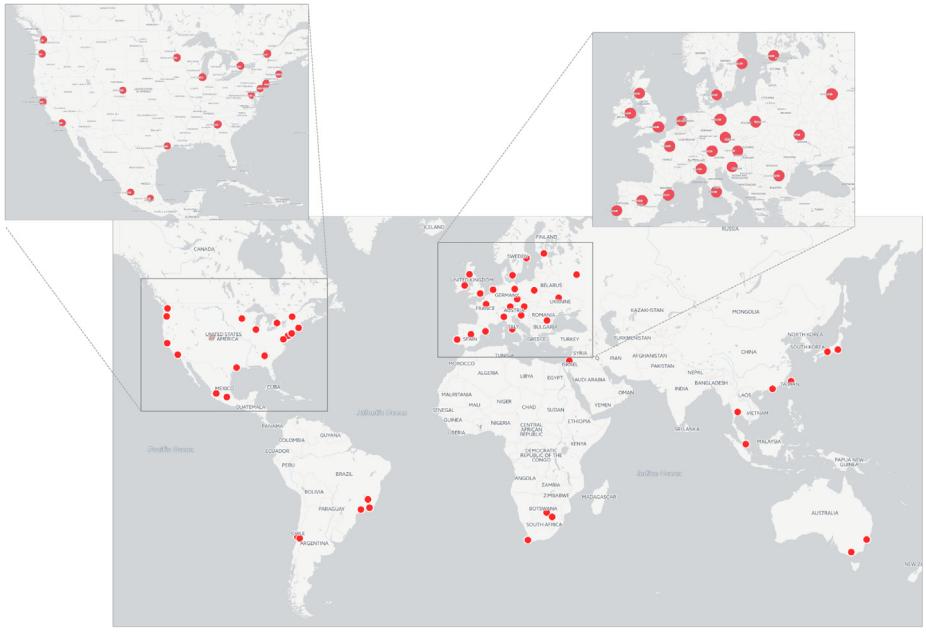

Fig. 2. Geographic distribution of the image data in the MIT Place Pulse dataset – data on 56 cities from all over the world are included in the dataset.

Fig. 3. Image samples from the MIT Place Pulse dataset with their perceptual score of the 6 dimensions.

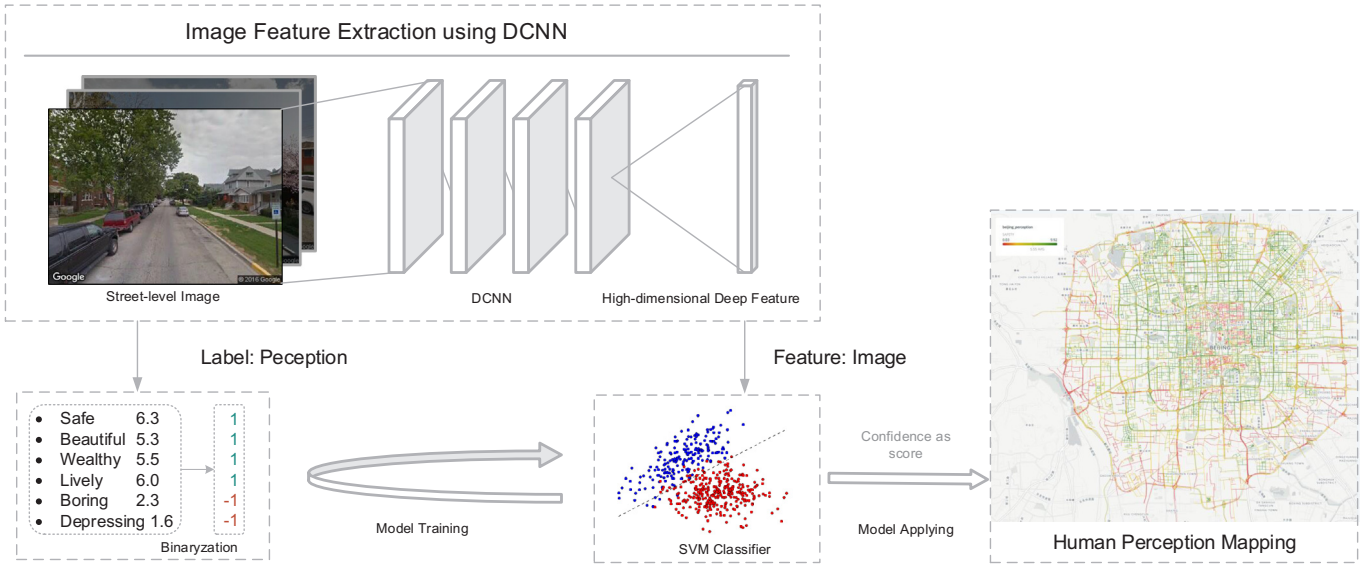

Fig. 4. An overview: predicting human perception of street-level imagery. First, we extract the image feature using DCNN and annotated each image with a binary label. Second, an SVM classifier is trained to predict the human perception of the street view images of a new urban region. Third, the spatial distribution of human perception is mapped.

individually. For each perceptual indicator, the whole dataset was simply sampled and split into a positive group and a negative group according to the samples' Q score.

To avoid introducing noise and error as much as possible, we selected representative positive/negative samples from the whole dataset to use for the training task. Specifically, for a specific perceptual indicator ν, we first calculated the mean value μ_ν, and standard deviation σ_ν of the dataset. We used a ratio variable δ to determine the threshold for sampling. Hence, if image i was selected with the score Q_i^ν, then its label y_i^ν could be represented as:

$$y_i^\nu = \begin{cases} -1 & \text{if} \quad Q_i^\nu < \mu_\nu - \delta\sigma_\nu \\ 1 & \text{if} \quad Q_i^\nu > \mu_\nu + \delta\sigma_\nu \end{cases} \tag{4}$$

Consequently, the two thresholds - $\mu_\nu - \delta\sigma_\nu$, and $\mu_\nu + \delta\sigma_\nu$ created a gap between the positive and negative samples, and the "noise" data lying in-between were removed. The data were annotated with the label "-1" and "1" for negative and positive samples respectively. The variable δ determined the bandwidth of the gap. In this study, an experiment was conducted with different δ to observe the model's performance, and Fig. 6 presents the specific number of samples used for the experiment.

We experiment with training on the high-dimensional deep features by using Radial Basis Function (RBF) kernel SVM (Joachims, 1998), which is a binary classifier with kernels seeking a linear boundary in higher dimensional space. For a typical Support Vector Classifier (SVC), the decision function can be represented by:

$$sgn\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right) \tag{5}$$

where $x_i \in \mathbb{R}^p$, $i = 1, ...c$, $andn$ are the training vectors, $y_i \in \{1, -1\}^n$ are the label classes, $K(x_i, x)$ denotes the Gaussian kernel, and α_i denote the parameters of the linear combination of the Gaussian kernels $K(x_i, x)$.

In this case, the Gaussian kernel - RBF, can be represented by:

$$K(x_i, x) = \exp\left(\frac{-||x-x'||}{2\sigma^2}\right) \quad (6)$$

4.2. Identifying sensitive visual elements

To determine the visual elements that may cause a place to be perceived as safe, lively, depressing, etc., we introduce a method to identify the visual elements of a place that are highly correlated with human perceptions.

The data used in this study were obtained from MIT Place Pulse dataset. Basically, we used the perceptual rating scores of each image from the calculation (Section 3.1). Meanwhile, to represent the street scene elements we employed semantic scene parsing techniques (Zhou et al., 2017) to calculate the area ratio of semantic objects in the scene. Semantic scene parsing is one of the key techniques used for scene understanding (Zhou et al., 2017), which aims at recognizing and segmenting object instances in one natural image. Given an input image, the model is able to predict one class label for each pixel. Enabled by the recently developed DCNN (LeCun et al., 2015), the state-of-the-art scene parsing model, PSPNet, has reached 79.70% pixel-wised accuracy in classifying 150 categories of objects (Zhao, Shi, Qi, Wang, & Jia, 2017), and has been employed in this study.

Fig. 5 is an overview of the multivariate regression analyses. The analyses were conducted separately for each of the six perceptual indicators. For one perceptual indicator, the perceptual scores were obtained from the MIT Place Pulse dataset, and the area ratio of each visual element in the image was calculated by counting the pixel numbers in the segmentation mask.

Multivariate regression analyses were employed to investigate the dependence between multiple variables. In this case, in consideration of the effects that bring by spatial autocorrelations, which causes measurements tobe clustered in related statistical units, we perform linear mixed model analyses, which add a random effect for the dependent variables to account for the correlations between data coming from the same cities. The mixed model can be represented as:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \epsilon \quad (7)$$

where $\boldsymbol{y}$ is a vector of observations; $\boldsymbol{\beta}$ is an unknown vector of fixed effects; $\boldsymbol{u}$ is an unknown vector of random effects; ϵ is an unknown vector of random errors; X and Z are known design matrices relating the observations $\boldsymbol{y}$ to $\boldsymbol{\beta}$ and $\boldsymbol{u}$, respectively.

Each of the 6 perceptual indicators was used as the dependent variable in 6 independent analysis individually, and the 150 object categories were treated as the predictors and also used as the fixed effects in the mixed model. Meanwhile, the cities from which the observations were obtained were used as random factors. The contribution of each object to a specific perceptual attribute was compared by observing the standardized coefficient of that object in the regression analysis.

5. Experiment and results

5.1. Perception prediction results

The RBF kernel SVM was trained and validated over fivefold cross-validation. The performance of the model with different δ values, namely 0.5, 0.7, 1.0, 1.2, 1.5, and 1.8 was evaluated. Fig. 6 shows the number of positive and negative samples used with different δ. To treat the imbalanced classes, weights were assigned to each sample in the training process, according to the size of splits. Moreover, we experimented with training tasks on six perceptual indicators: safe, lively, beautiful, wealthy, depressing, and boring. The results are shown in Fig. 6.

As we can see from the Fig. 6, high and reliable accuracy were achieved in predicting the six perceptual indicators. The accuracies of the different indicators varied. For instance, the accuracy of safe, beautiful, and wealth were slightly higher than that of depressing, boring and lively. This result might be caused by variances in how people understand these concept; their knowledge might tend to be relatively consistent with “what is the beautiful scene” but inconsistent with “what is the depressing scene”. Another reason might be insufficient data collection for the latter three dimensions. In addition, the average accuracy decreased as the bandwidth of the sample gap narrowed (smaller δ), indicating that human preference was comparatively unstable for normal scenes.

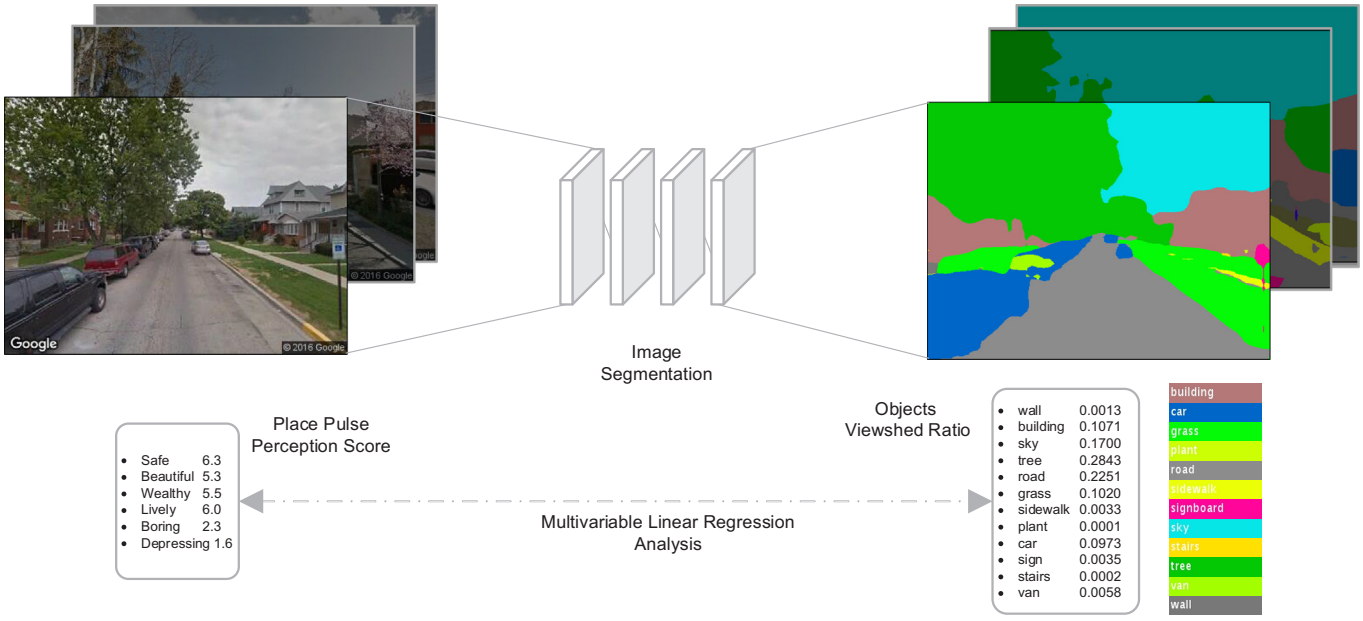

Fig. 5. An overview: multivariate linear regression analyses between the perceptual scores and the presence of a visual element. Image samples are selected from the Place Pulse dataset with perceptual scores. The presence of a visual element is calculated from the image using an image semantic segmentation model.

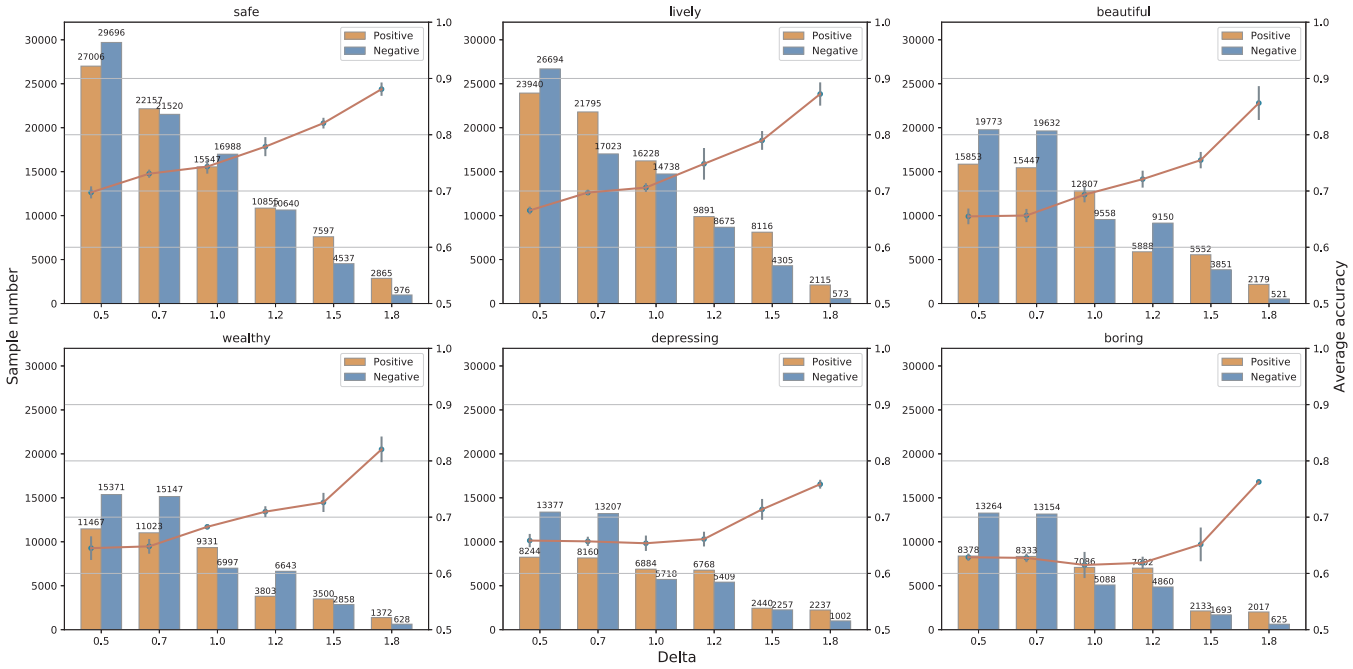

Fig. 6. The sample size and the corresponding average accuracy in the experiment are shown. The vertical bars show the positive and negative samples used in the training task with different values of δ. As δ increases, fewer samples were selected. The red curves indicate the average accuracy with different sizes of training samples. The error bar was calculated from 5-fold cross-validation.

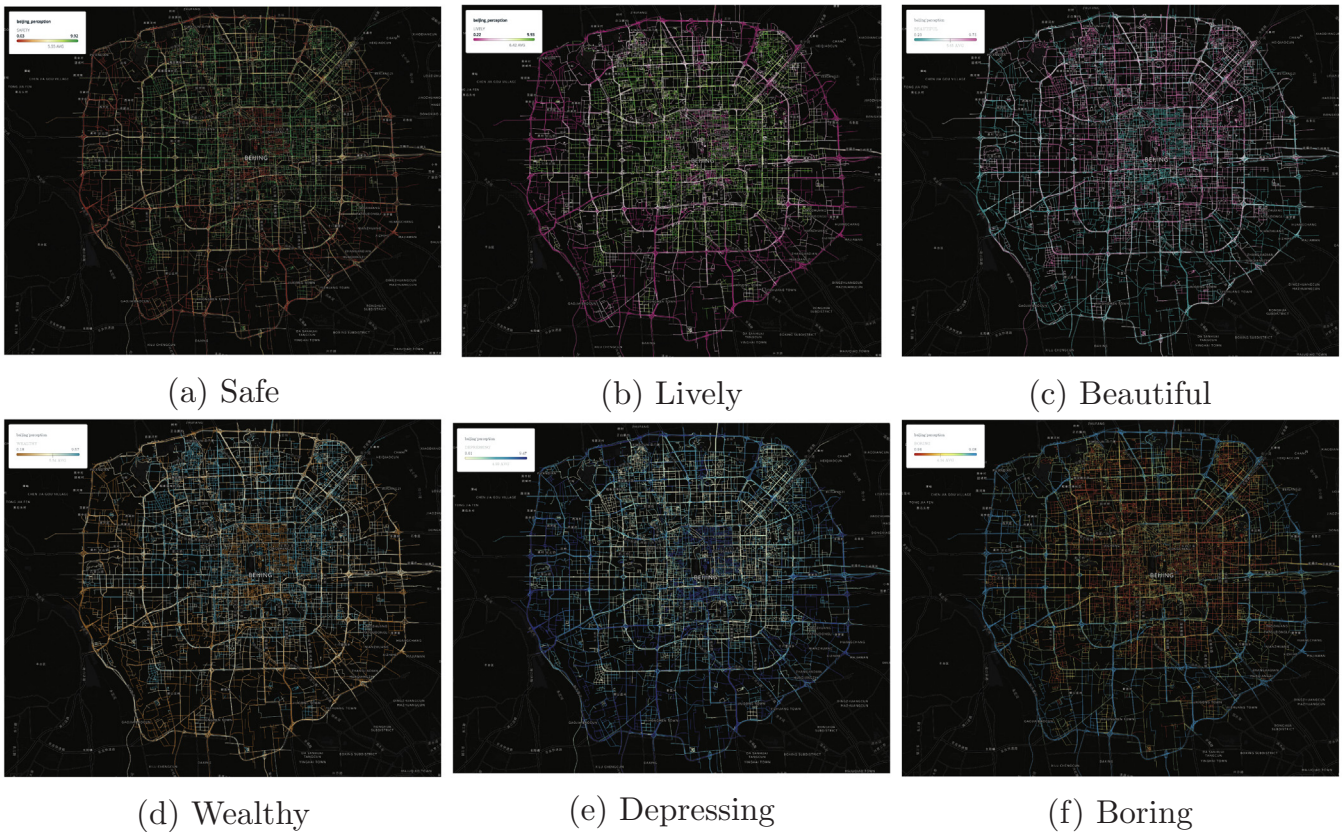

Fig. 7. Mapping the human perception of Beijing using 6 perceptual indicators.

5.2. Perception mapping in a new urban region

The pre-trained perception predictors were used to estimate the perceptual distributions in a new region. However, since the predictor was binary, the estimation result was discrete and in the form of -1 and 1. To quantify the perception and obtain a continuous value, we took the positive confidence of the SVM model, which indicates the probability of one sample to the positive label, as the value of the

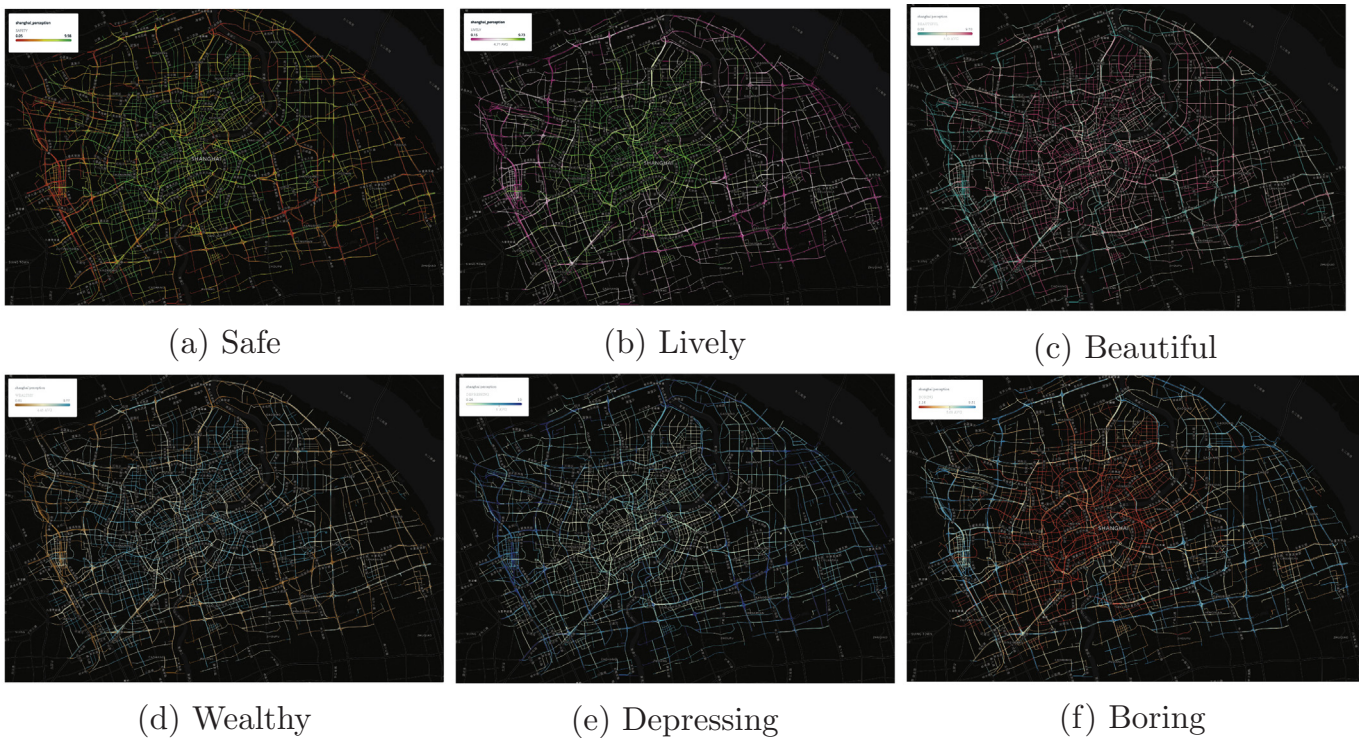

Fig. 8. Mapping the human perception of Shanghai using 6 perceptual indicators.

Fig. 9. Image samples from Shanghai that were predicted with high safe scores (left) and low safe scores (right).

perception.

The perceptual scores of all images for Beijing and Shanghai were collected in this work and calculated by the pre-trained model. We then used the street as the visualization unit and spatially joined the image points as well as their perception scores to its street by averaging. Fig. 7 and Fig. 8 show the perceptual mapping of the two cities, which can be considered the “City Perception Map”.

In general, from the figure, we can conclude that the downtown areas are more “safe” and “lively” than the surrounding suburbs. Similarly, mid-level roads are more “safe” and “lively” than the ring road and highway. However, there are several groups of street networks inside Beijing’s third ring road that is predicted to be unsafe, which is due to the large number of old houses and narrow roads that have fallen into disrepair around these historical sites. Moreover, we notice that the short and small streets in the dense street networks tend to be more lively. This result may align with Jacobs (Jacobs, 1992), who suggested

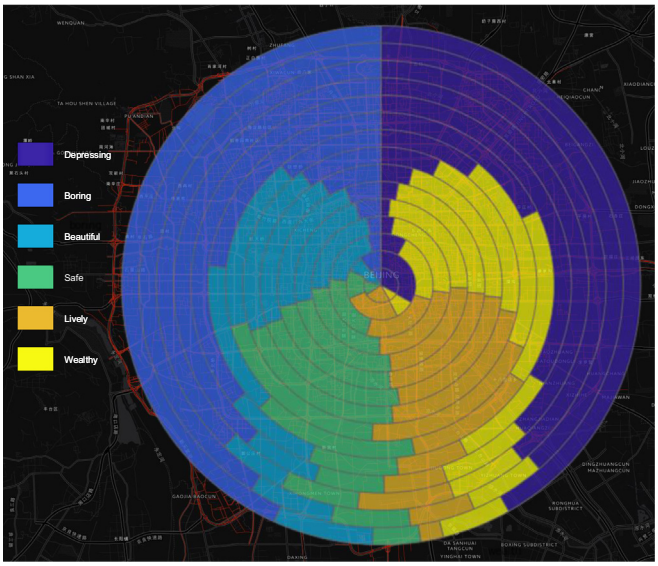

Fig. 10. Spatial Changes of the six perceptual indicators from the inner city to the outer city of Beijing using concentric rings. The Beijing area was divided into 16 concentric rings of a 1 km incremental radius starting at the center of Tiananmen Square. For each concentric zone, the perceptual scores of the images were aggregated, averaged and normalized. Note that 16 the concentric zones indicated the actual spatial location, while the sub-zones of each concentric zone only denoted the relative ratio of the 6 perceptual indicators.

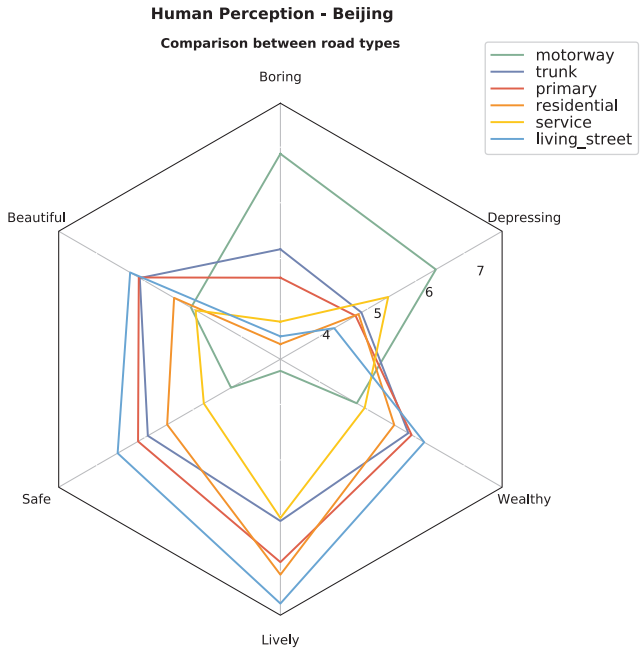

Fig. 11. Perceptual indicators of different road types in Beijing.

that blocks should be designed to be short, to increase path options and diversity. People's perceptions actually align with the actual functional area of the city. For example, there are several communities distributed around the area between the fourth and fifth ring road, which is actually a growing residential area. Fig. 9 presents several street view images sampled from Shanghai. The samples that were predicted to have high safe scores are shown on the left, and samples with low safe scores are shown on the right. The prediction results match up with our intuition that the well-attended streets with green plants and vehicles tend to be more safe, while the areas with awnings, scruffy fences, and mud roads are considered to be unsafe.

Perceptual spatial distribution of Beijing. Inspired by the preliminary results of the perception spatial distributions, we were interested in the spatial change in the perceptual indicators from the inner city to the outer city. Taking Beijing as a case study, we took Tiananmen Square as the center of circle, and divided the research area into 16 concentric rings of a 1 km incremental radius. By aggregating, averaging and normalizing the perceptual scores of the images in each concentric zone, we were able to observe the spatial change in the perceptual indicators. As shown in Fig. 10, generally, from the inner city to the outer city, the indicators "beautiful", "safe", "lively", and "wealthy" first increased and then decreased, and the indicators "depressing " and "boring" showed the opposite pattern. In particular, the area with the most "positive" perceptions was the 5th ring, which corresponds to the area around the third ring road of Beijing.

Perceptions of different road types. From Fig. 7 and Fig. 8 we noticed that the perception scores were highly related to the type of road. To understand the human perceptions of different types of roads, we aggregated and averaged the perceptual scores according to road type.

Based on the descriptions obtained from the OpenStreetMap, the streets in the road network of Beijing were classified into 6 major categories: motorway, trunk, primary, residential, service, and living street. The description of each category is shown in Table 2. As we can see from the results (Fig. 11), "motorways", which were significantly different from the other road types, received high values for "boring", and "depressing" and low values for the four positive perceptions. The perceptions for "trunk" and "primary" roads were similar, while the latter one was more "lively" and less "boring". Of all the road types, "living streets" had the highest value for the four positive perceptions and the lowest value for the two negative perceptions.

Correlation analyses among the perceptions. By comparing the distribution of the human perception across the 6 indicators, we also noticed that the geo-spatial patterns were intuitively similar among some of the perceptual indicators. Consequently, a crossover Pearson correlation analysis of the 6 indicators was conducted. Fig. 12 shows the Pearson correlation coefficients with data from Beijing and Shanghai. Generally, we found that some of the pairs of the indicators were highly correlated, such as "beautiful – wealthy" and "depressing - safe", and some pairs like "beautiful - boring" were relatively independent. More specifically, the connections among these indicators varied between Beijing and Shanghai. For instance, the correlation of "wealth - depressing" was strong in Beijing but comparatively low in Shanghai. We believe that the inconsistency is caused by the different characteristics of cityscapes of the two cities, and more in-depth studies are needed to

Table 2
Road type and description.

Type	Num.	Description (from Open Street Map)
Motorway	835	A restricted access major divided highway, normally with 2 or more running lanes plus emergency hard shoulder.
Trunk	1191	The most important roads in a country's system that aren't motorways.
Primary	1555	The next most important roads in a country's system.
Residential	3029	Roads which serve as an access to housing, without function of connecting settlements. Often lined with housing.
Service	1666	For access roads to, or within an industrial estate, camp site, business park, car park etc.
Living street	226	For living streets, which are residential streets where pedestrians have legal priority over cars, speeds are kept very low and where children are allowed to play on the street.

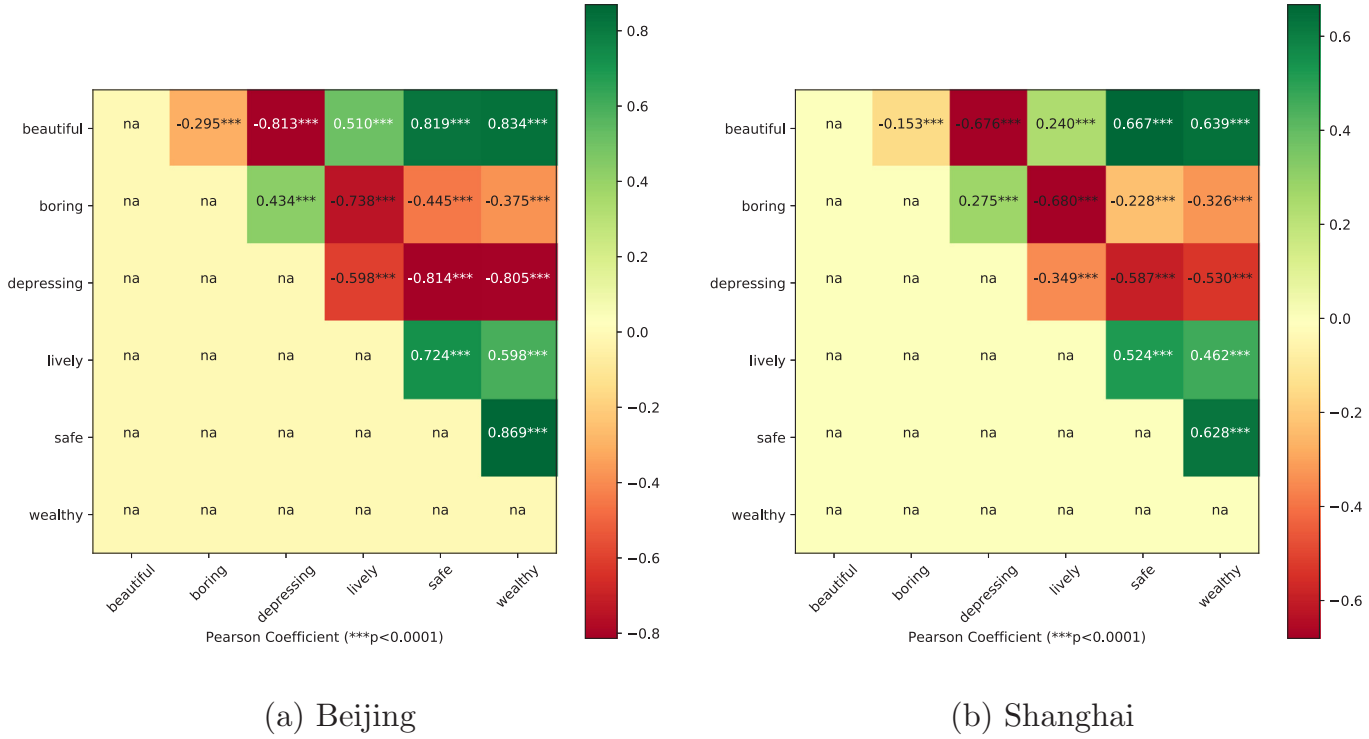

Fig. 12. Pearson correlation coefficients among the 6 perceptual indicators using data from Beijing (a) and Shanghai (b).

test this hypothesis in the future.

5.3. Factor identification results

In Fig. 13, we present the results of the multivariate regression analyses between the perceptual indicators and the presence of visual elements, where the top 10 objects that positively (red bar) or negatively (blue bar) contributed to each perceptual indicator are ranked and listed. The length of the bar indicates the value of the betas - standardized coefficients and the * implies the significance level.

As shown in Fig. 13, overall, we found that the objects that highly contribute to the perceptual indicators varied. For instance, the area ratios of “cars”, “sidewalks”, and “roads” were positively correlated with the “lively” score. This result exactly aligns with Jacobs’s proposition (Jacobs, 1992) to improve the liveliness of a city. In contrast, “beautiful”,“wealthy”, and “depressing” were more sensitive to the areas of “trees”, “grass” and ‘flora’. Moreover, we noticed that a “wall” was a negative element in almost all of the six dimensions (here we refer to “positive” for “depressing” and “boring”), especially for “safe”, which is counter-intuitive to some extent. This finding corresponds to the philosophy of contemporary urban planning, which has argued that walls may lead to blocked views, decreased sunshine and the build-up of pollution (Wong, Nichol, & Ng, 2011) and thus should be reduced. This phenomenon was coined the “wall effect”. Our result provides evidence of the “wall effect” from the perspective of human perception.

The fact that urban greenery brings a sense of peacefulness and quietness has been discussed by Ashihara in his theory about the art of landscape in streets (Ashihara, 1983) and by Rachel and Stephen Kaplan in their theory about “restorative environments” (Kaplan & Kaplan, 1989). Consistent with these theories, the “greenery” and “natural” objects were highly correlated with all the perceptual indicators. From another perspective, our results showed that the positive contributors of “beautiful” came from natural elements, where objects such as buildings were excluded. Such a relationship also agrees with Olmsted’s Philosophy of embedding the eco-system into urban infrastructure. The objects “building” and “minibike” were positively correlated with “lively”, which is also consistent with the goals of new urbanism (Beveridge & Rocheleau, 1995).

Regarding the sense of safety, the disorder of the physical setting of a place, such as the filthiness of streets caused by litter, graffiti, vandalism, and poorly maintained buildings, reduces the perception of feeling “safe” (Skogan & Maxfield, 1981; Taylor, Gottfredson, & Brower, 1984; Wilson & Kelling, 1982). People may choose to take a different route if they perceive a neighborhood to be unsafe (Short, 1984). Similarly, Perkins points out that personalization of property can make the street environment a safer look, as can the presence of streetlights, block watch signs, yard decorations and private plantings (Perkins, Meeks, & Taylor, 1992). In this study, streetlights and traffic signs were not identified as predictors of feeling “safe”, because of their small volume in the images used in this study. However, similar things such as cars, plants, and houses were to be found positively correlated with the sense of safety.

6. Discussion

The physical setting of a place and its perceptions impact the behavior and health of its dwellers. For more than a century, a wide variety of fields have discussed the importance of urban physical appearance and the visual factors that may contribute to human perceptions. The contributions of this work are as follows: first, this study used an approach to determine the perceptions of a place for a large-scale urban region using big data of street-level imagery. Second, this study sought to define the connections between the physical setting of a place and the human perceptions of the place quantitatively.

The large-scale mapping of human perceptions provides a macroscopic perspective for observing the whole cities and urban regions. We can see many opportunities to apply this method practically. For example, in pedestrian navigation applications, the perceptual map of a city is able to suggest a path that will be more comfortable instead of time-consuming, giving users a special walking experience. In terms of theoretical implications, we believe street-level imagery offers potential opportunities for place formalization, for instance, enriching place semantics with human perceptions, which will help researchers understand the underlying urban heterogeneity patterns and reveal the

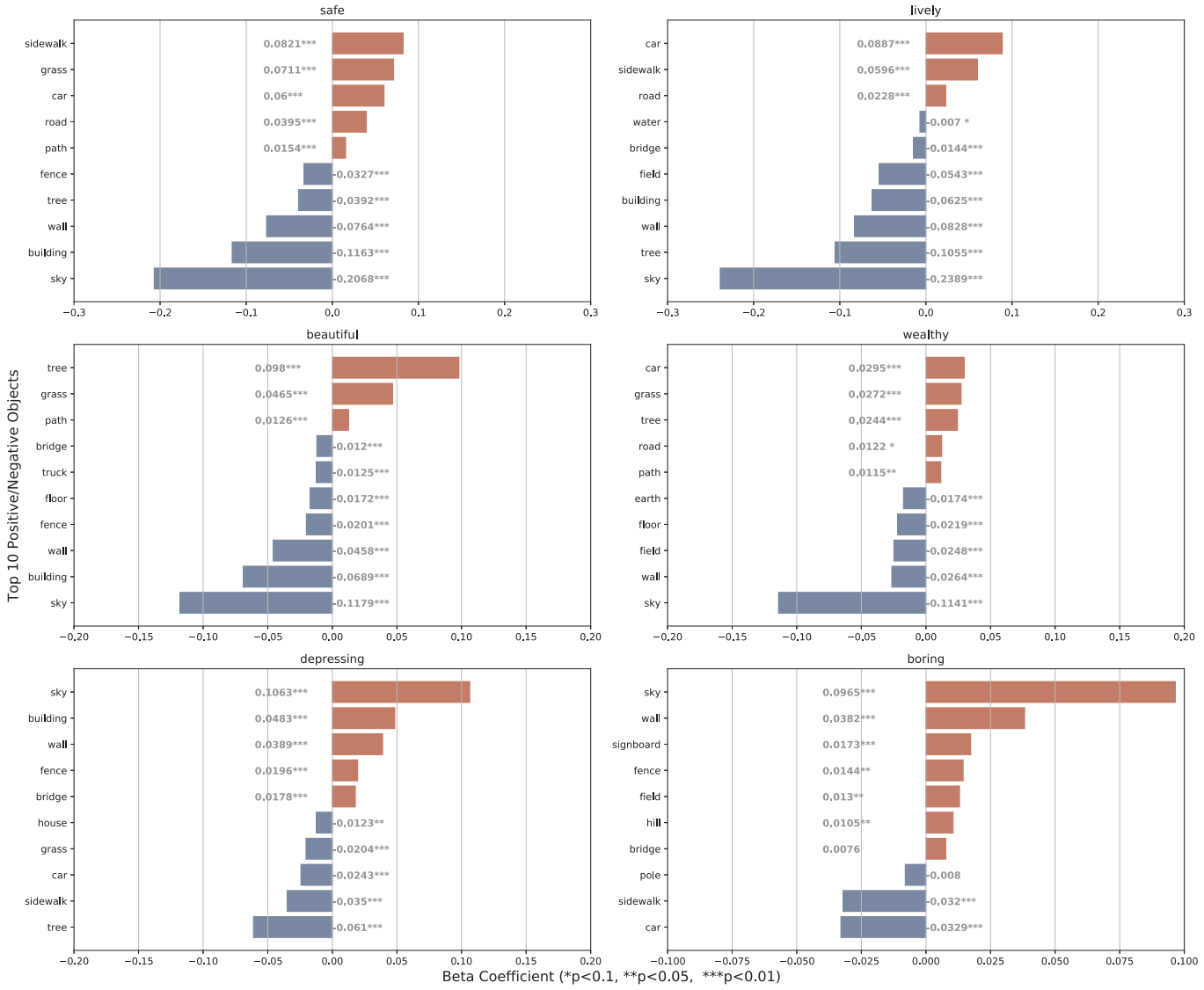

Fig. 13. The results of the multivariate regression analysis between the scene elements and the perception scores. For each pair, the pixel number of a particular object category and the perception score along a specific dimension are given. The top 10 objects that positively/negatively contributed to each of the 6 perception types are shown.

impacts of urban function. In addition, perceptual mappings in a time-series will indicate urban changes, which will also support theories connecting the physical setting of place and social-economic variables. For example, Naik, Kominers, Raskar, Glaeser, and Hidalgo (2017) found that neighborhoods with a certain demographic property are more likely to be physically improved.

What visual elements impact human perceptions? In Section 5.3, we conduct a multivariate regression analysis to identify the presence of visual elements that correlated with human perceptions in terms of six dimensions, namely, safe, lively, beautiful, wealthy, depressing and boring. On one hand, the visual elements identified in this study, for instance, the contribution of walls, green plants, vehicles, and man-made features to specific perceptual indicators, will directly support the theory and practice of urban design. On the other hand, the potential benefit of these results lies in creating computer-generated scenes based on derived criteria using data-driven rendering techniques, for example, generative adversarial nets (GAN) (Goodfellow et al., 2014). These techniques have a great potential to generate urban scenes that are perceived safe, lively and non-depressing with presenting a certain proportion of visual elements and to further inform urban designers.

The limitations of this work also deserve to be discussed and to be paid more attention to in future works. First, in applying our model to map the human perceptions of new urban regions, the bias that was caused by the visual variances in the landscape of the different cities should be considered. Technically, this problem can be formulated as adapting data distribution between different domains. To validate the performance of the prediction phase, future works may need to collect more samples from the region, conduct an evaluation experiment and test the model with the test set.

Second, this is a preliminary study in terms of connecting visual elements in street view images with human perceptions. In particular, it is challenging to encode all the variances in the human perceptions of a place. As implied in previous studies (Quercia et al., 2014), human perceptions extend beyond the visual, and experiencing a place is not about observing singular viewpoints or looking at a specific visual object, experiencing a place is more about the culture, history, activities, and interactions with the surroundings over time, which is not easily be represented by visual images. Future studies to model the place sentiment could by possibly extended to investigate social and humanistic factors using data such as point of interest (POI), human mobility, and demographics of the neighborhood.

In addition, the image data we used in this study were street view

images, which were taken along urban streets. Street is one major place where human activities take place, but not all of them. Areas in street blocks such as parks, alleyways, vacant lots, campus, etc. also contribute powerfully to people's perceptions. In future studies, images from social media that cover more wide urban spaces should also be considered and extended to the analysis.

7. Conclusion

The physical setting of a place and its perceptions impact the behavior and health of its dwellers; however, measuring human perceptions of a place in large-scale urban regions has been challenging due to the lack of both quantitative data and appropriate methods to deal with the data. In this work, we propose a quantitative approach to measure human perceptions of a large-scale urban environment in an automatic and efficient way by incorporating street-level imagery to represent places and a deep learning method to understand the high-level information of the images. The study provides a tool to better understand human perceptions of the built environment. Second, we conduct statistical analyses to identify the visual elements that highly impact human perceptions. Specifically, we correlate the presence of visual elements with human perception scores to determine what kind of physical setting has an impact on the sense of place. We have identified that a wall is a negative element in an urban scene, and natural elements always induce positive perceptions, which is compatible with the literature in related fields.

The results of this work support urban design theories and practices and illustrate the value of employing machine learning methods to understand how people perceive the physical setting of places. This study also demonstrates the value of street-level imagery in place formalization. Place semantics may be enriched in terms of the human sense, helping researchers understand the underlying urban structure and reveal the impacts of urban function.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2017YFB0503602, the National Natural Science Foundation of China under Grant 41671378 and 41625003, and the Hong Kong Research Grants Council under GRF Grant No. 14606715. Authors wish to thank Dr. Yaoli Wang for her helpful advice and comments on this paper, and to thank the MIT SENSEable City Lab Consortium for supporting this research.

References

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., ... Weaver, J. (2010). Google street view: Capturing the world at street level. *Computer, 43*(6), 32–38.

Arase, Y., Xie, X., Hara, T., & Nishio, S. (2010). Mining people's trips from large scale geo-tagged photos. *Proceedings of the 18th ACM international conference on multimedia* (pp. 133–142). ACM.

Ashihara, Y. (1983). *The aesthetic townscape.* The MIT press.

Beveridge, C. E., & Rocheleau, P. (1995). *Frederick law olmsted.* Rizzoli International Publications.

Crandall, D. J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the world's photos. *Proceedings of the 18th international conference on World Wide Web* (pp. 761–770). .

Cresswell, T. (1992). *In place-out of place: Geography, ideology, and transgression, Vol. 1.* University of Minnesota Press.

Cresswell, T. (2014). *Place: an introduction.* John Wiley & Sons.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. *European conference on computer vision* (pp. 288–301). Springer.

Datta, R., Li, J., & Wang, J. Z. (2008). Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. *Proceedings of the IEEE conference on image processing* (pp. 105–108). IEEE.

Deza, A., & Parikh, D. (2015). Understanding image virality. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 1818–1826.

Dhar, S., Ordonez, V., & Berg, T. L. (2011). High level describable attributes for predicting aesthetics and interestingness. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1657–1664). IEEE.

Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A. (2012). What makes Paris look like Paris? *ACM Transactions on Graphics, 31*(4).

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

Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2016). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics,* 249–256.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Goodchild, M. F. (2011). *Formalizing place in geographic information systems. In Communities, neighborhoods, and health.* Springer pp. 21–33.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems,* 2672–2680.

Hays, J., & Efros, A. A. (2015). Large-scale image geolocalization. *Multimodal location estimation of videos and images* (pp. 41–62). Cham: Springer.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *IEEE international conference on computer vision* (pp. 2980–2988). IEEE.

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

Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill: a bayesian skill rating system. *Advances in Neural Information Processing Systems,* 569–576.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., ... Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 145–152). IEEE.

Jacobs, J. (1992). The death and life of great American cities.

Jae Lee, Y., Efros, A. A., & Hebert, M. (2013). Style-aware mid-level representation for discovering visual connections in space and time. *Proceedings of the IEEE international conference on computer vision,* 1857–1864.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* (pp. 137–142). .

Kaplan, R., & Kaplan, S. (1989). The experience of nature: A psychological perspective. *CUP Archive*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Less, E. L., McKee, P., Toomey, T., Nelson, T., Erickson, D., Xiong, S., & Jones-Webb, R. (2015). Matching study areas using google street view: A new application for an emerging technology. *Evaluation and Program Planning, 53*, 72–79.

Liu, L., Zhou, B., Zhao, J., & Ryan, B. D. (2016). C-IMAGE: city cognitive mapping through geo-tagged photos. *GeoJournal, 81*(6), 817–861.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*(3), 512–530.

Lynch, K. (1960). *The image of the city, Vol. 11.* MIT press.

Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. *Proceedings of the 18th ACM international conference on multimedia* (pp. 83–92). ACM.

Mallgrave, H. F. (2010). *The architect's brain: Neuroscience, creativity, and architecture.* John Wiley & Sons.

Michael, R. (2005). Online visual landscape assessment using internet survey techniques. *Trends in online landscape architecture: Proceedings at Anhalt University of Applied Sciences* (pp. 121). .

Mišić, B., Fatima, Z., Askren, M. K., Buschkuehl, M., Churchill, N., Cimprich, B., & Korostil, M. (2014). The functional connectivity landscape of the human brain. *PLoS One, 9*(10), e111007.

Montello, D. R., Goodchild, M. F., Gottsegen, J., & Fohl, P. (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation, 3*(2–3), 185–204.

Morison, B. (2002). *On location: Aristotle's concept of place.* Oxford University Press on Demand.

Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2017). Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences, 114*(29), 7571–7576.

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

Nasar, J. L. (1997). *The evaluative image of the city.* Sage Publications.

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

Perkins, D. D., Meeks, J. W., & Taylor, R. B. (1992). The physical environment of street blocks and resident perceptions of crime and disorder: Implications for theory and measurement. *Journal of Environmental Psychology, 12*(1), 21–34.

Porzi, L., Rota Bulò, S., Lepri, B., & Ricci, E. (2015). Predicting and understanding urban perception with convolutional neural networks. *Proceedings of the 23rd ACM international conference on multimedia* (pp. 139–148). ACM.

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

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems,* 91–99.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

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

Schroeder, H. W., & Anderson, L. M. (1984). Perception of personal safety in urban recreation sites. *Journal of Leisure Research, 16*(2), 178.

Short, J. R. (1984). *An introduction to urban geography.* London: Routledge & Kegan Paul.

Skogan, W. G., & Maxfield, M. G. (1981). Coping with crime: Individual and neighborhood reactions.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems,* 3104–3112.

Taylor, R. B., Gottfredson, S. D., & Brower, S. (1984). Block crime and fear: Defensible space, local social ties, and territorial functioning. *Journal of Research in Crime and Delinquency, 21*(4), 303–331.

Tuan, Y.-F. (1977). *Space and place: The perspective of experience.* University of Minnesota Press.

Ulrich, R. S. (1983). *Aesthetic and affective response to natural environment. In Behavior and the natural environment.* Springer pp. 85–125.

Ulrich, R. S. (1984). View through a window may influence recovery from surgery. *Science, 224*(4647), 420–421.

Valtchanov, D., & Ellard, C. G. (2015). Cognitive and affective responses to natural scenes: effects of low level visual properties on preference, cognitive load and eye-movements. *Journal of Environmental Psychology, 43*, 184–195.

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. *Atlantic Monthly, 249*(3), 29–38.

Wong, M. S., Nichol, J., & Ng, E. (2011). A study of the wall effect caused by proliferation of high-rise buildings using gis techniques. *Landscape and Urban Planning, 102*(4), 245–253.

Zhang, F., Zhang, D., Liu, Y., & Lin, H. (2018). Representing place locales using scene elements. *Computers, Environment and Urban Systems, 71*, 153–164.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. *Proceedings of IEEE conference on computer vision and pattern recognition,* 2881–2890.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ADE20K dataset. *Proceedings of the IEEE conference on computer vision and pattern recognition*.