Received: 15 November 2023 | Revised: 19 March 2024 | Accepted: 22 April 2024

DOI: 10.1111/tgis.13172

RESEARCH ARTICLE

Transactions in GIS WILEY

Urban perception by using eye movement data on street view images

Nai Yang[1] | Zhitao Deng[1] | Fangtai Hu[1] | Yi Chao[1] | Lin Wan[2] | Qingfeng Guan[1] | Zhiwei Wei[3,4]

[1]School of Geography and Information Engineering, China University of Geosciences, Wuhan, 430078, China

[2]School of Computer Science, China University of Geosciences, Wuhan, 430078, China

[3]Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100830, China

[4]The Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100830, China

Correspondence
Lin Wan, School of Computer Science, China University of Geosciences, Wuhan 430078, China.
Email: wanlin@cug.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 42171438; Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing, Grant/Award Number: KLIGIP-2023-B06; Open Research Fund of Key Laboratory of Engineering Geophysical Prospecting and Detection of Chinese Geophysical Society, Grant/Award Number: CJ2021IC04; University-Industry Collaborative Education Program, Grant/Award Number: 220600273083107

Abstract

Understanding the spatial distribution patterns of urban perception and analyzing the correlation between human emotional perception and street composition elements are important for accurately understanding how people interact with the urban environment, urban planning, and urban management. Previous studies on urban perception using street view data have not fully considered the actual level of attention to different visual elements when browsing street view images. In this article, we use eye tracking technology to collect eye movement data and subjective perception evaluation data when people browse street view images, and analyze the correlation between the time to first fixation, duration of first fixation, and fixation frequency of different visual elements and the six perceptual outcomes of wealthy, safe, lively, beautiful, boring, and depressing. Furthermore, this article integrates eye movement data with street view semantic data and introduces a novel method for predicting urban perception using a machine learning algorithm. The proposed method outperforms a comparative model that solely relies on semantic data, exhibiting higher accuracy in perception prediction. Additionally, the study presents a perceptual mapping of the prediction results, providing a visual representation of the predicted urban perception outcomes. As vision is the primary perceptual channel, this study achieves a more

objective and scientifically reliable urban perception, which is of reference value for the study of physical and mental health due to the urban physical environment.

1 | INTRODUCTION

The visual appearance of cities is a central factor in shaping human emotions and perceptions of the surrounding urban environment (Kelling & Coles, 1997; Porzi et al., 2015). The visual quality of an urban space can significantly impact the psychological state of its inhabitants, which in turn influences people's emotions, behaviors and decision-making (Lindal & Hartig, 2013). Thus, people's emotional perceptions and evaluations of urban spaces directly impact their perception and experience of cities (Wolch et al., 2014). Understanding the people's emotional responses and evaluations of the city enables the planning and design of the city's spatial layout at the macro level, and the management and regulation of the city's spatial resources at the micro level.

In the real world, vision is the primary mode of acquiring spatial information (Fabrikant et al., 2010). Meanwhile, the human brain and vision are the most direct and effective way to perceive the surrounding environment, and are inherently superior in global understanding and recognition of natural scenes (Greene & Oliva, 2009a, 2009b). The process of human emotional cognition is subject to attentional biases toward certain stimuli that are often associated with emotions (Krassanakis & Cybulski, 2019). With the advancement of biological and psychological instruments, eye tracking devices have become a valuable tool for measuring a person's visual attention, providing rich sources of perceptual information, such as where, when, how long, and in what order certain information is viewed in or about space (Kiefer et al., 2017). Additionally, eye tracking has been shown to be a reliable and effective method for measuring emotional responses to visual stimuli, providing valuable information for the design and management of urban spaces (Wu et al., 2022).

In this study, a combination of eye tracking and deep learning techniques is employed to effectively capture the subjective visual perceptual features of observers and the semantic information associated with each pixel in natural images (Zhao et al., 2017). The collected data are then efficiently analyzed to extract relevant feature and perceptual information embedded within street scenes. Finally, mathematical learning models, such as random forests, are utilized to establish correlations between specific urban physical environments and the psychological experiences of urban residents. This quantitative calculation allows for the determination of the level of urban perception. In general, the combination of visual information and environmental features to analyze the perception and evaluation of urban space makes urban perception research more objective, scientific, and accurate.

2 | RELATED WORKS

2.1 | Research on urban perception

Traditional studies often use on-site assessments or field observations to collect data on urban environment characteristics (Dadvand et al., 2016; Herzog, 1992). However, this method is time-consuming and labor-intensive to obtain urban residents' perception level, limiting its scalability for large-scale studies. With the development of high-performance computing and deep learning technologies, researchers have utilized residents' perceptions to quantitatively assess urban environments in the fields of urban planning, sociology, and geo-information science. For example, Salesses et al. (2013) proposed the concept of computational human perception of cities and created the Place Pulse dataset to map how different areas of a city are perceived.

Subsequently, the MIT Media Lab's Place Pulse 2.0 collects urban appearance ratings by presenting participants with two street view images and asking them to choose the one that appears more X (e.g., safe, beautiful, depressing, lively, wealthy, boring) (Dubey et al., 2016). Building upon this, Zhang et al. (2018) further advanced the field by employing a data-driven machine learning approach to measure human perception of urban areas. Meanwhile, Yao et al. (2019) proposed a human-computer adversarial scoring framework to assess urban perceptions quickly and economically in Chinese cities. The proposed method provided perception estimates with less than 10% error, and demonstrated its feasibility in facilitating the derivation of local urban perceptions. More recently, Wang et al. (2022) evaluated the perception of street scenes across six dimensions using a deep learning and spatial syntax approach, with the top 20% of the highest-rated street scenes considered to be high-quality street spaces and the top 20% of the lowest-rated street scenes considered to be low-quality street spaces. The study also investigated the relationship between different perceptual dimensions and street components in a high accessibility context.

All of the above studies directly used semantic segmentation data of street view images to train urban perception models, which solely relied on the size of the segmentation area to determine feature influence on perception. However, this approach overlooked the impact of semantic content on human perception, potentially undervaluing features with smaller areas and lower occurrence frequencies. To address this, we incorporated eye movement data capturing visual-level perception and cognitive-level human interest. This integration aimed to train a perception model that eliminates semantic imbalance, resulting in a more scientific, accurate, and objective representation of human perception.

2.2 | Application of eye movement tracking in spatial cognition

Eye tracking is the process that tracks the movement of the eyes to know exactly where a person is looking and for how long (Klaib et al., 2021). Eye tracking systems measure the position, movement and pupil size of the eyes at specific times to detect areas of interest to the user (Fairbairn & Hepburn, 2023). In the area of spatial cognition, eye movement data were used to investigate the recognition of activities on a map, including searching for points of interest and planning routes (Kiefer et al., 2013). Mobile eye tracking technology has also been used to evaluate different interface designs for indoor pedestrian navigation systems displaying landmarks, with researchers assessing the merits of maps with different levels of detail by comparing subjects' wayfinding performance (Ohm et al., 2016). With the development of artificial intelligence technologies, researchers can now combine eye tracking and machine learning techniques to explore human behavior and cognitive processes (Zemblys et al., 2018). For example, Liao et al. (2018) conducted a realistic walking navigation experiment to extract features from eye movement data in a realistic environment for a pedestrian navigation scenario. They used five eye movement features to classify five common navigation tasks by a random forest classifier. However, Noland et al. (2016) noted the lack of research that combines eye tracking technology to study the relationship between gaze preference and human perception of urban space. To address this gap, they used noninvasive eye tracking technology to quantitatively examine areas of interest and gaze data. They also overlaid these data with qualitative evaluations of the image. Oki and Kizawa (2021) developed a deep learning model to estimate attractiveness ratings of street scene images in a dense log cabin area. They explored the link between participants' gaze patterns and their attractiveness ratings for the street images. The researchers employed Grad-CAM to visualize the model's decision structure and identify street components contributing to attractiveness assessment.

In summary, eye tracking is commonly used in spatial cognition research to evaluate visual interfaces and infer human behavior. Because of the cognitive significance of eye movement metrics, the coupling of street view images and machine learning methods can efficiently, objectively, and accurately calculate the perception level of city dwellers on a large scale, helping to design a more realistic and complete emotional map of the city.

3 | EYE MOVEMENT EXPERIMENT BASED ON STREET VIEW IMAGES

3.1 | Data source

The study area selected for this research is the eight major administrative districts of Wuhan, a central city in central China, including Wuchang, Hongshan, Jiang'an, Jianghan, Qiaokou, Hanyang, Qingshan and Dongxihu (Figure 1). The street view images were obtained from Tencent Maps (https://map.qq.com/) and the road network data were obtained from OpenStreetMap.org. In this study, sampling points were extracted every 100 m along the major road networks in the central city, and for each sampling point, the street view images were collected in four horizontal views of 0°, 90°, 180°, and 270° (Figure 2). Helbich et al. (2019) showed that the horizontal perceptual view was also more consistent with the urban perception of city residents. A total of 99,428 street view images were collected from 24,857 sampling points in Wuhan.

To uncover the impact of environmental visual factors on human perception of activities and social interactions (Wu et al., 2022), we conducted pixel-level semantic segmentation of street view images. This allowed us to determine the level of visual attention subjects directed to different environmental elements in the location. In this study, a deep learning Fully Convolutional Network based on the ADE_20K dataset training proposed by Yao et al. (2019) was used to semantically segment the street view images of the main city of Wuhan into 151 feature types, including unknown features.

3.2 | Design and implementation of eye movement experiments

To obtain information about the distribution of people's attention to different features during image viewing, we selected 2040 out of nearly 100,000 Street View images as the experimental eye movement dataset and the rest as the estimate dataset, with a uniform image size of 480×300 pixels. Figure 3 illustrates the distribution of

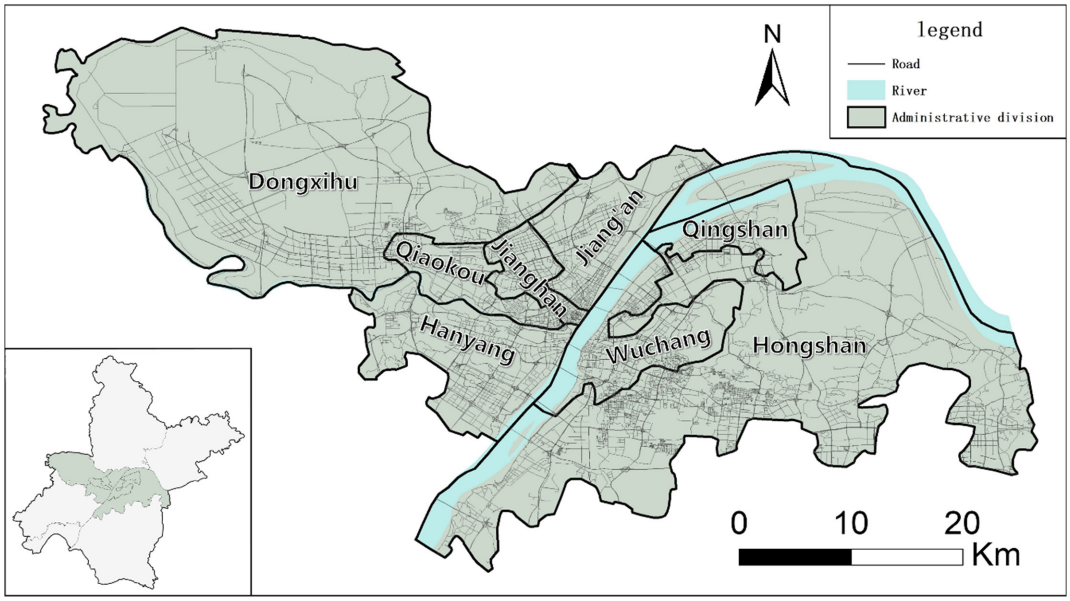

FIGURE 1 Overview of the study area: eight central urban areas of Wuhan, Hubei Province, with the thin black line showing the main roads in the study area. The lower left corner shows the relative position of the study area (in dark) in the administrative division of Wuhan City.

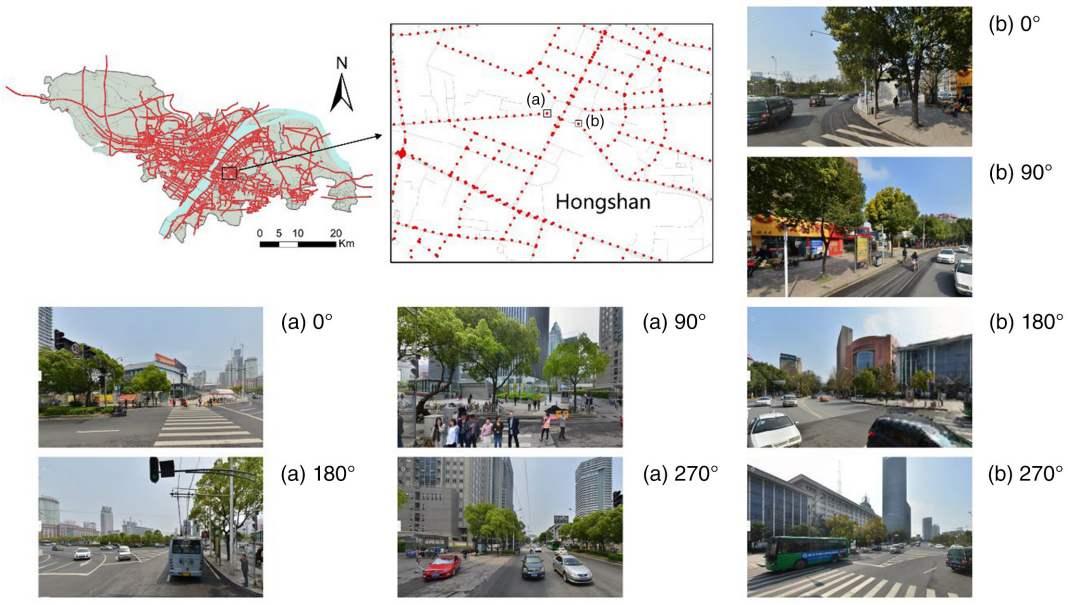

FIGURE 2 Examples of street view images: (a) intersection of Democracy Road and Hongshan Square, (b) intersection of Bayi Road and Hongshan Square.

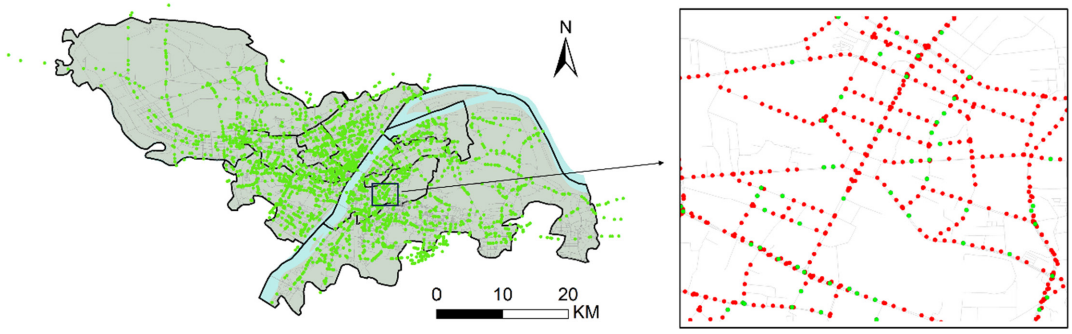

FIGURE 3 Spatial distribution of sampling points: green is the experimental dataset sampling points; red is the estimated dataset sampling points.

coordinates for the sampling points within the dataset. By carefully screening the sampling points, we aimed to maintain a uniform distribution of points throughout the study area. This approach enabled us to capture a broad representation of streetscapes across the entire city, ensuring that our dataset encompasses a diverse range of urban environments. The experimental dataset of street view images was sorted by sample point coordinates into 34 groups of 60 images each.

During the experiment, each subject was asked to view one street image that appeared randomly on the screen, and the process of viewing the image was controlled by the subject. The sampling rate of the eye tracker was 120 Hz, the tracking distance was 50–90 cm, and the subjects were asked to keep a distance of about 70 cm from the eye tracker. Based on the experience of Dubey et al. (2016)'s research, six perceptual types were selected: wealthy, safe, lively, beautiful, boring, and depressing. To quantify the perceptual values of each type, their value ranges were uniformly set to [0,100], with 0 being the lowest and 100 being the highest (Yao et al., 2019). Eye movement experiments and perceptual ratings were conducted simultaneously to ensure the reliability of eye movement acquisition. The subjects' gaze remained fixed on the street scene image displayed on the screen

TABLE 1 Descriptive statistics of the questionnaire scores.

	Min	Max	Mean	S.D.
Wealthy	17	89	46.93	11.264
Safety	9	82	40.24	8.056
Lively	3	91	39.77	13.702
Beautiful	2	94	41.67	11.862
Boring	34	83	60.61	6.683
Depressing	22	89	54.62	8.685

throughout the experiment. Simultaneously, a trained assistant collected perceptual data by asking questions and recording the subjects' scores. Human ethics has been approved by the Research Ethics Board of our institution. Descriptive statistics of the questionnaire scores are shown in Table 1.

Overall, 67 current undergraduate, graduate students, teachers, government employees, and businessmen in Wuhan were used as subjects. All subjects were free of eye diseases (e.g., high myopia, low color vision, or astigmatism) and had visual acuity or corrected visual acuity of 1.0, including 34 males and 33 females, aged between 18 and 60. Each subject was randomly assigned to view between 5 and 10 sets of images (selected from the 34 groups described above). As a result, the street images in the experimental dataset were observed 12 times by various subjects. All subjects had eye movement sampling rates exceeding 70% on each street view image. Additionally, we employed leave-one-out cross-validation methodology in our study. This involved randomly selecting 60 images and systematically excluding the eye movement data of one subject from each image in a cyclical manner. By comparing the gaze distributions before and after the exclusion, we obtained an average correlation coefficient of 96.37%. This approach was adopted to minimize significant differences in eye movement behavior among individuals and enhance the ecological validity of the eye movement data. In the end, 2040 street images with valid eye movement data and perception scores were recorded, and the average fixation duration of each image was 28.63 s.

3.3 | Eye movement feature extraction for experimental dataset

Base on sections 3.1 and 3.2, we were able to identify the environmental elements that each gaze point attended to by analyzing segmented semantic labels and gaze point locations. To extract eye movement metrics, we classified each semantic category as an area of interest for visual attention. In this study, four eye movement metrics were selected for each area of interest (AOI): total duration of fixation (TDF), number of fixations (NF), time to first fixation (TFF), and duration of first fixation (DFF). To eliminate the influence of the browsing time of each street view image on the eye movement metrics, the four types of eye movement metrics obtained are normalized by using the percentage transformation. This ensures that the values of each type of eye movement metrics of different features are between 0 and 1. As shown in formula 1, metric_{ij} represents the i-th class eye movement metric value of the j-th class visual element in a street view image, where i ranges from 1 to 4 and corresponds to the four types of eye movement metrics mentioned above. $\sum_{j=1}^{n} \mathrm{metric}_{ij}$ represents the sum of the i-th type eye movement metric values of all visual elements in a street view image. And Metric_{ij} represents the normalized result of the i-th class eye movement metric value of the j-th class visual element in a street view image.

The study also introduces TFF, DFF, and fixation frequency (FF) to investigate the most important visual elements that influence perceptual scores. The FF in AOI is a metrics that takes into account both the TDF and the NF in AOI. It represents the number of gaze points in the AOI per unit of time. A higher value indicates more interest in the target, reflecting the importance of the evaluation area indicators (Dong et al., 2019;

Guo et al., 2017). The ratio of Count_i, the number of all fixations in the i-th region of interest, to Duration_i, the duration of the region of interest, is defined as the Frequency_i of the region of interest, as shown in formula 2. Table 2 reflects the descriptive statistics of semantic segmentation and eye movement metrics for the top 15 visual elements.

$$\text{Metric}_{ij} = \frac{\text{metric}_{ij}}{\sum_{j=1}^{n} \text{metric}_{ij}} \tag{1}$$

$$\text{Frequency}_i = \frac{\text{Count}_i}{\text{Duration}_i} \tag{2}$$

4 | CORRELATION ANALYSIS OF PERCEPTUAL DIMENSIONS AND EYE MOVEMENT FEATURES

4.1 | Results of correlation analysis

Linear regression analyses were conducted on TFF, DFF, and FF, as well as perception scores of wealthy, safe, lively, beautiful, boring, and depressing for each visual element. The degree of influence between the eye movement factor and perception were described using bate coefficients, with significance levels indicated by asterisks (*). Our goal was to determine the impact of visual elements on human gaze and how they correlated with human perception scores.

4.1.1 | The correlation between time to first fixation in AOI and perception

Figure 4 shows the results of a multiple linear regression of TFF in areas of interest and perceptual dimensions. The top 10 visual elements with the highest standardized coefficients were selected for each distinct emotion, and they have been ranked and presented accordingly. Cross-sectionally, for car, building, sidewalk, person, and traffic light are all positively correlated with wealthy, safe, and lively perceptions, while TFF for field and ground are all negatively correlated with wealthy, safe, and lively perceptions. In addition, TFF for grass and lake are positively correlated with beautiful perception and negatively correlated with boring and depressing perceptions. TFF of ground and column are negatively correlated with beautiful perception and negatively correlated with boring and depressing perceptions. Longitudinally, we can observe that across the six dimensions of perception, the TFF of different visual elements is mostly positively correlated with the four positive emotions of wealthy, safe lively, and beautiful perceptions, and negatively correlated with the two negative emotions of boring and depressing perceptions. In psychological cognition, TFF is usually analyzed together with DFF (Garza et al., 2016; Yang et al., 2023), this phenomenon is analyzed in the following Section 4.1.2.

4.1.2 | The correlation between duration of first fixation in AOI and perception

Figure 5 shows the results of a multiple linear regression of the DFF in AOI and the perceptual dimensions. The results indicate that wealthy, safe, and lively perceptions are positively correlated with the DFF of car, sidewalk, traffic light and building, and negatively correlated with the DFF of ground, grass, and field. Similarly, beautiful perception was positively correlated with the DFF of grass and lake, and negatively correlated with the DFF of ceiling, column, and ground. In contrast, boring and depressing perceptions were negatively correlated with the duration of the first fixation of grass, lake, or plant, and positively correlated with the duration of the first fixation

TABLE 2 The descriptive statistics of semantic segmentation and eye movement metrics for the top 15 visual elements.

	Semantic segmentation				Time to first fixation (TFF)				Duration of first fixation (DFF)				Fixation frequency (FF)			
	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.
Building	0.000	0.901	0.179	0.175	0.000	1.000	0.081	0.168	0.000	1.000	0.173	0.171	0.000	1.000	0.156	0.140
Tree	0.000	0.924	0.152	0.162	0.000	1.000	0.110	0.189	0.000	1.000	0.157	0.162	0.000	1.000	0.144	0.131
Road	0.000	0.424	0.179	0.104	0.000	1.000	0.114	0.177	0.000	1.000	0.099	0.113	0.000	1.000	0.107	0.109
Car	0.000	0.284	0.029	0.042	0.000	1.000	0.086	0.160	0.000	1.000	0.094	0.135	0.000	1.000	0.077	0.103
Sky	0.000	0.708	0.265	0.166	0.000	1.000	0.112	0.201	0.000	1.000	0.088	0.122	0.000	1.000	0.113	0.130
Ground	0.000	0.433	0.015	0.042	0.000	1.000	0.029	0.108	0.000	1.000	0.022	0.073	0.000	1.000	0.023	0.070
Field	0.000	0.404	0.002	0.016	0.000	1.000	0.005	0.048	0.000	1.000	0.005	0.036	0.000	1.000	0.004	0.033
Sign	0.000	0.362	0.004	0.013	0.000	1.000	0.025	0.099	0.000	0.972	0.020	0.066	0.000	0.670	0.021	0.063
Grass	0.000	0.344	0.013	0.031	0.000	0.991	0.035	0.116	0.000	0.832	0.025	0.073	0.000	0.551	0.026	0.070
Sidewalk	0.000	0.356	0.031	0.046	0.000	1.000	0.059	0.143	0.000	0.808	0.037	0.083	0.000	0.800	0.042	0.086
Plant	0.000	0.613	0.016	0.039	0.000	1.000	0.035	0.114	0.000	0.778	0.024	0.069	0.000	0.731	0.027	0.072
Wall	0.000	0.637	0.018	0.053	0.000	0.999	0.033	0.110	0.000	0.768	0.027	0.075	0.000	0.936	0.029	0.074
Trade Name	0.000	0.069	0.001	0.003	0.000	1.000	0.006	0.056	0.000	0.724	0.007	0.048	0.000	0.632	0.006	0.041
Fence	0.000	0.227	0.012	0.024	0.000	1.000	0.046	0.129	0.000	0.655	0.035	0.080	0.000	0.782	0.037	0.080
Lake	0.000	0.205	0.001	0.008	0.000	0.609	0.002	0.024	0.000	0.557	0.002	0.021	0.000	0.291	0.002	0.018
Ceiling	0.000	0.503	0.008	0.046	0.000	0.960	0.004	0.046	0.000	0.523	0.003	0.024	0.000	0.523	0.003	0.028

Note: All values are between 0 and 1.

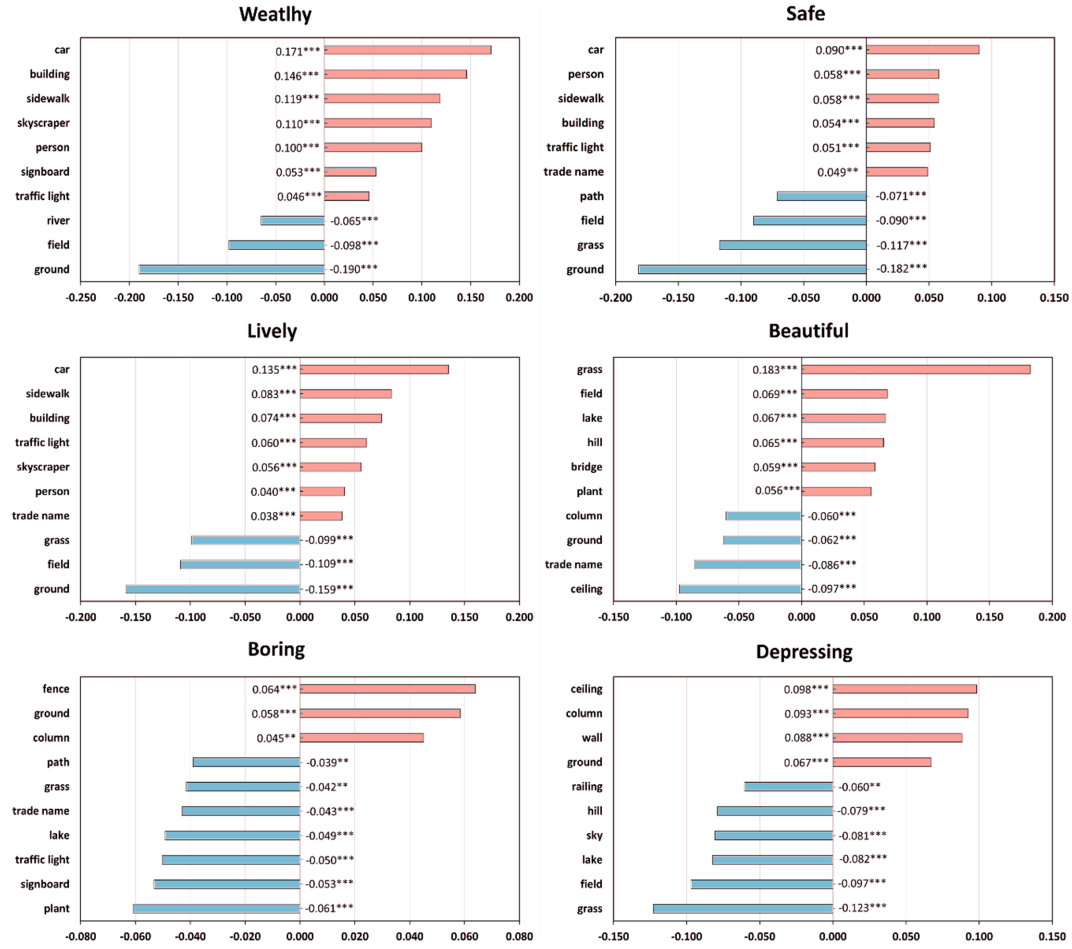

FIGURE 4 The results of a linear regression analysis of T for the visual elements against the six perceptual dimensions show the top 10 objects with the greatest positive/negative contribution for each perception type. Beta coefficient (**$p < 0.01$,***$p < 0.001$)

of ceiling, column, and ground. These findings are very consistent with the above findings on the TFF of corresponding visual elements.

Section 4.1.1 shows that the TFF of different visual elements are mostly positive in positive emotion and negative in negative emotion. The longer the TFF means that the target is less conspicuous (Cavicchio et al., 2014), suggesting that these visual elements with lesser salience within the street view images tend to evoke more positive emotions. The longer the DFF, the stronger the attractiveness of the target (Garza et al., 2016), indicating that the attractiveness of these visual elements plays an important role in the perception of positive emotion. Consequently, individuals exhibit a tendency to seek visually appealing elements that evoke positive emotions when exploring street view images, irrespective of their prominence. In psychology, there is a cognitive bias called positive Affect Bias, which refers to people's preference and priority for positive emotional information in information processing and attention allocation (Tamir & Robinson, 2007). This bias can influence the perception of visual stimuli and the allocation of attention. In addition, German philosopher and art theorist Ernst H. Gombrich's theory of the 'aesthetic response' also suggests that esthetic stimuli trigger positive emotions and pleasure (Gombrich, 1995). Therefore, people tend to look for visual elements that are optimistic, harmonious, and lively when looking at street view images in order to experience positive emotions and pleasure.

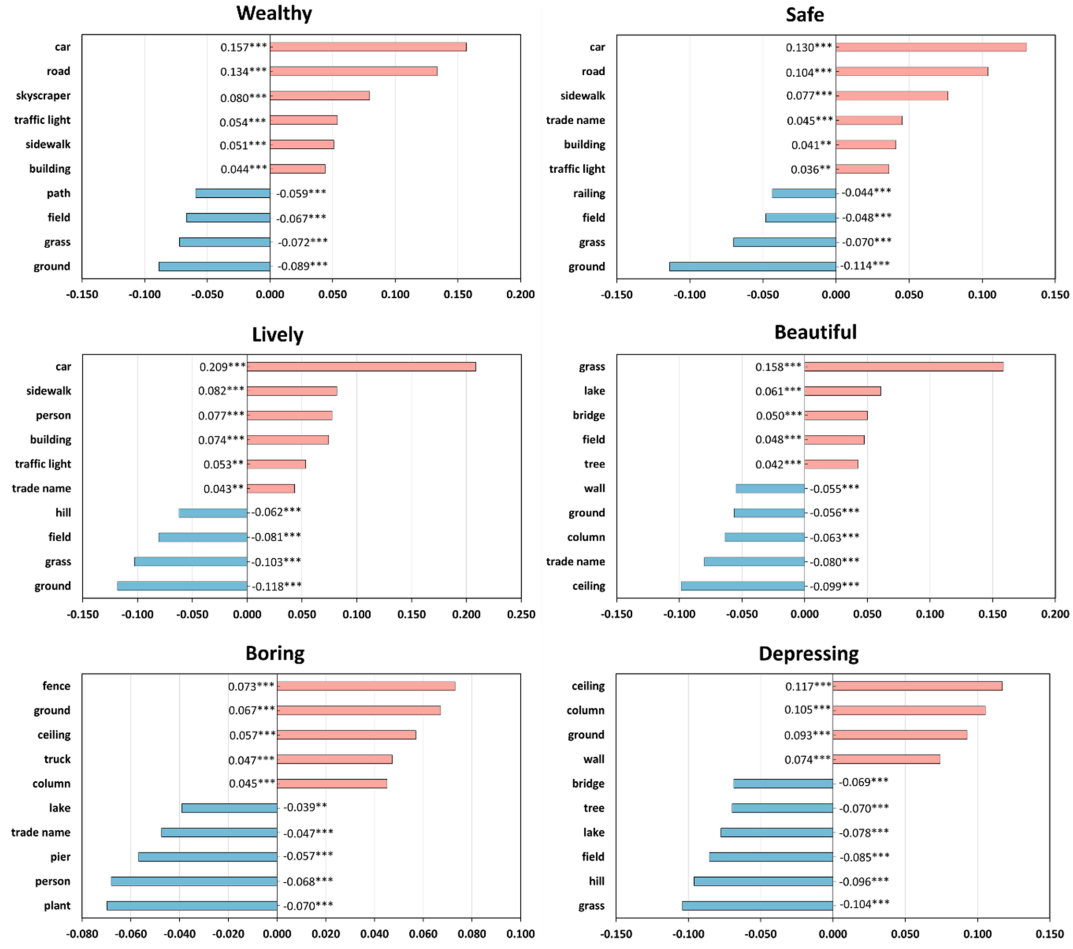

FIGURE 5 The results of a linear regression analysis of the DFF for the visual elements against the six perceptual dimensions show the top 10 objects with the greatest positive/negative contribution for each perceptual type. Beta coefficient (**$p < 0.01$,***$p < 0.001$)

It is noteworthy that the correlation between TFF and the DFF with perception for some visual elements is not entirely consistent. In visual cognition, TFF is the searching stage of perception, while the DFF means entering the discovery stage of memory search and information processing (Shao, 2019; Zheng, 2020), and human perception scores are the result of information processing, which shows that the correlation between DFF and perception is more important.

4.1.3 | The correlation between fixation frequency in AOI and perception

Figure 6 shows the results of a multiple linear regression of FF in AOI on the perceptual dimensions. The 10 visual elements with the highest contribution to emotional perception are also selected for ranking, and other visual elements with lower salience were not displayed. In the same cross-sectional comparison, the FF of car, road, skyscraper, and sidewalk are positively correlated with wealthy, safe, and lively perceptions, while the FF for railing is negatively correlated with wealthy and lively perceptions, and these types of features also represent the level of urban infrastructure development and economic development in common sense. The FF

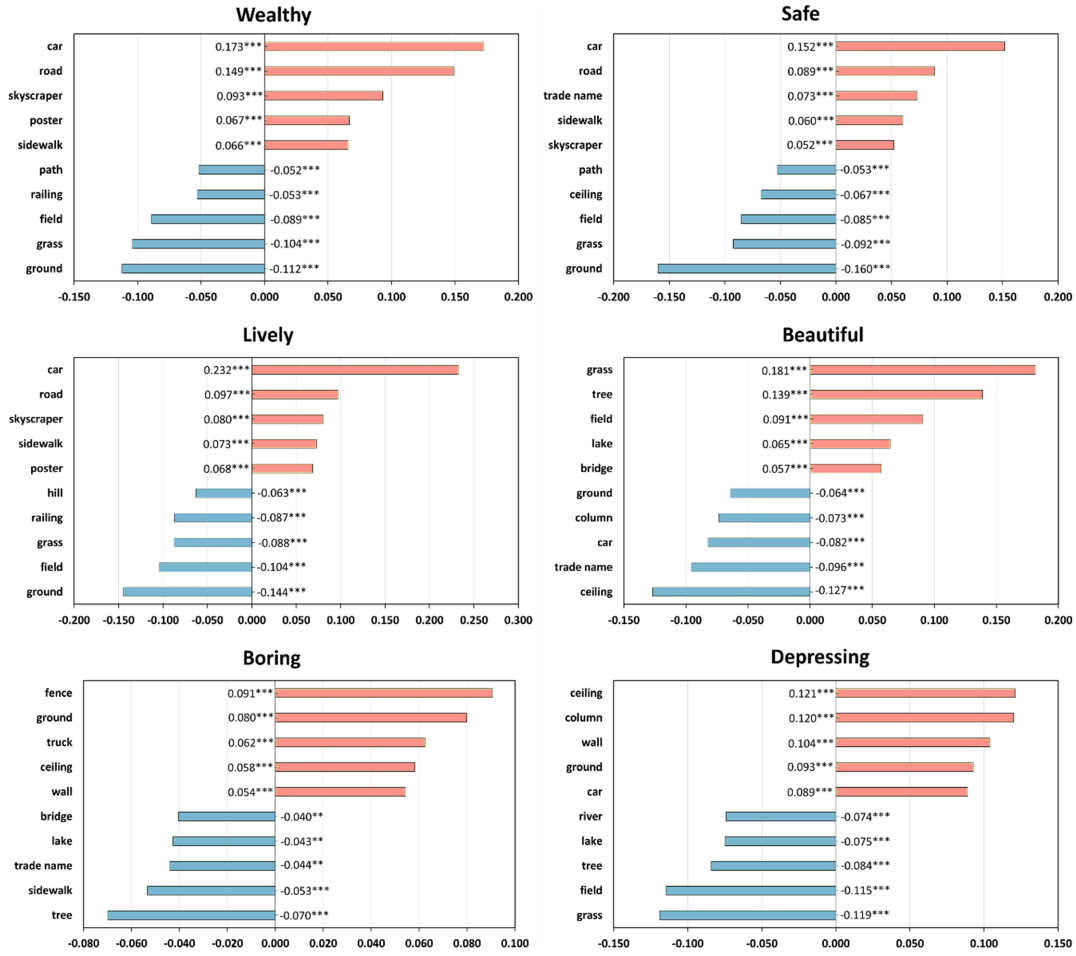

FIGURE 6 The results of a linear regression analysis of the FF for the visual elements against the six perceptual dimensions show the top 10 objects with the greatest positive/negative contribution for each perception type. Beta coefficient ($^{**}p < 0.01$,$^{***}p < 0.001$)

of ground was negatively correlated with wealthy, safe, lively, and beautiful perceptions, but positively correlated with boring and depressing perceptions, in line with the findings by Wang et al. (2022) in the Binjiang district of Hangzhou, China. Barren, unvegetated ground does not have esthetic value in the city and is also a reflection of the poor infrastructure of the community (Zhang et al., 2022). Longitudinally, wealthy and lively perceptions are negatively correlated with the FF of natural landscapes such as grass and field. The beautiful perception is positively correlated with the FF of vegetation landscape, such as grass, tree and field, and negatively correlated with the FF of trade name. This is corroborated by Pieters and Wedel (2004) in their study on the attention capture of images and text ads. They pointed out that images in advertising had a significant advantage over copy in attracting attention, and that people also had a preference for images such as posters over text-based trade names. Another interesting phenomenon is that beautiful perception is positively correlated with the FF of lake and bridge, while boring and depressing perceptions are negatively correlated with the FF of lake, bridge, or river. The inverse correlation observed between negative perceptions and visual elements associated with water suggests that a greater interest in water bodies and related visual elements strengthens the inhibitory effect on negative perceptions. Wuhan, renowned as the "lake capital" of China, stands at the forefront among the country's major cities in terms of water area. The region boasts an intricate

network of rivers and extensive water conservancy facilities, earning its reputation as the "bridge capital" of China (Chen et al., 2022; Lin & Liu, 2012). This shows that the hydrological and hydraulic features of Wuhan city, such as river and lake landscapes and bridge projects, have positive guiding effects on people's mental health. This conclusion was better confirmed in the study by Chen et al. (2022).

4.2 | The research findings

To visually depict the influence of the eye movement metrics of the various visual elements on each perception, Table 3 summarizes the correlations between the eye movement metrics of the 10 main visual elements and each perceptual dimension, obtained according to Figures 4–6. In the table, "+" indicates a positive correlation, "–" denotes negative correlation, "*" denotes correlation strength, and "/" denotes no significant correlation. Figure 7 shows some examples of segmentation for specific visual elements mentioned in Section 4.1, such as ceiling, column, pier, etc.

From the table, it can be observed that certain perceptual dimensions, the three different eye movement metrics of visual elements have exactly the same effect. For instance, TFF, DFF, and FF of land have negative correlations for the perception of wealthy, safe, lively and beautiful, and positive correlations for both boring and depressing perceptions. This suggests that heightened attractiveness and increased interest in land are associated with perceptions of boredom, depression, as well as a lack of wealth, safe, liveliness, and beauty. TFF, DFF, and FF of the car were all positively correlated on wealthy, safe, and beautiful perceptions. This suggests that the stronger the attractiveness and higher the level of interest of the car, the more it makes people perceive wealthy, safe, and beautiful. TFF, DFF, and FF of the hill are negatively correlated in terms of safe, lively, and depressing perceptions, and positively correlated in terms of beautiful perception. Similarly, these findings suggest that the stronger the attractiveness and higher the degree of interest of the hill, the more beautiful and unsafe, unlively, and non-depressing it is perceived to be.

In contrast, different eye movement metrics of some visual elements have different effects on perceptions, such as FF of trade name has no significant correlation with wealthy and lively perceptions, but it is TFF and the DFF have positive correlations with these two perceptions. This suggests that the more attractive the trade name is in the street view images, the wealthier and livelier it is perceived to be. Similarly, the DFF of the ceiling has no significant correlation with wealthy and lively perceptions, but its FF has a negative correlation with these two perceptions. This suggests that the more attractive the ceiling is in the street view images, the more they perceive wealthy and lively. The findings indicate that relying solely on a single eye movement metric is inadequate for studying urban perception. A comprehensive approach incorporating multiple eye movement metrics and diverse cognitive perspectives is necessary to ensure the scientific rigor of research results.

5 | URBAN PERCEPTION COMBINED EYE TRACK DATA AND SEMANTICS

The above studies show that eye movement metrics are closely related to urban perception. Therefore, in this study, eye movement metrics are added to the prediction model as important features that affect perception when exploring quantitative prediction methods for urban perception. Figure 8 shows the technical roadmap of urban perception prediction, which consists of three main steps: (1) constructing an urban perception model using eye movement data on the experimental dataset; (2) regressing eye movement data on the predicted dataset using semantic segmentation data; (3) performing perception prediction on the predicted dataset and plotting the perception results.

TABLE 3 Correlation between eye movement metrics of major visual elements and each perceptual dimension.

	Wealthy	Safety	Lively	Beautiful	Boring	Depressing
Car						
TFF	(+)***	(+)***	(+)***	(+)***	(−)***	(−)***
DFF	(+)***	(+)***	(+)***	(+)**	/	/
FF	(+)***	(+)***	(+)***	(+)**	/	(+)***
Grass						
TFF	(−)*	(−)***	(−)***	(+)***	(−)**	(−)***
DFF	(−)***	(−)***	(−)***	(+)***	/	(−)***
FF	(−)***	(−)***	(−)***	(+)***	/	(−)***
Field						
TFF	(−)***	(−)***	(−)***	(+)***	/	(−)***
DFF	(−)***	(−)***	(−)***	(+)***	/	(−)***
FF	(−)***	(−)***	(−)***	(+)***	/	(−)***
Sidewalk						
TFF	(+)***	(+)***	(+)***	/	/	(+)***
DFF	(+)***	(+)***	(+)***	(−)**	/	(+)***
FF	(+)***	(+)***	(+)***	(−)***	(−)***	(+)***
Ground						
TFF	(−)***	(−)***	(−)***	(−)***	(+)***	(+)***
DFF	(−)***	(−)***	(−)***	(−)***	(+)***	(+)***
FF	(−)***	(−)***	(−)***	(−)***	(+)***	(+)***
Wall						
TFF	/	/	/	(−)**	(+)**	(+)***
DFF	/	/	/	(−)***	/	(+)***
FF	(−)*	/	/	/	(+)***	(+)***
Trade name						
TFF	(+)*	(+)***	(+)***	(−)***	(−)***	(+)***
DFF	(+)*	(+)***	(+)**	(−)***	(−)***	/
FF	/	(+)***	/	(−)***	(−)**	/
Lake						
TFF	/	/	/	(+)***	(−)***	(−)***
DFF	/	/	/	(+)***	(−)**	(−)***
FF	/	/	/	(+)***	(−)***	(−)***
Ceiling						
TFF	(−)**	(−)*	(−)**	(−)***	/	(+)***
DFF	/	(−)**	/	(−)***	(+)***	(+)***
FF	(−)**	(−)**	(−)***	(−)***	(+)***	(+)***
Hill						
TFF	(−)**	(−)**	(−)**	(+)***	/	(−)***
DFF	(−)***	(−)*	(−)***	(+)***	/	(−)***
FF	(−)***	(−)**	(−)***	(+)***	/	(−)***

Abbreviations: DFF, the duration of first fixation; FF, fixation frequency; TFF, time to first fixation.
*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

FIGURE 7 Examples of segmentation instances for some specific visual elements mentioned in Section 4.1.

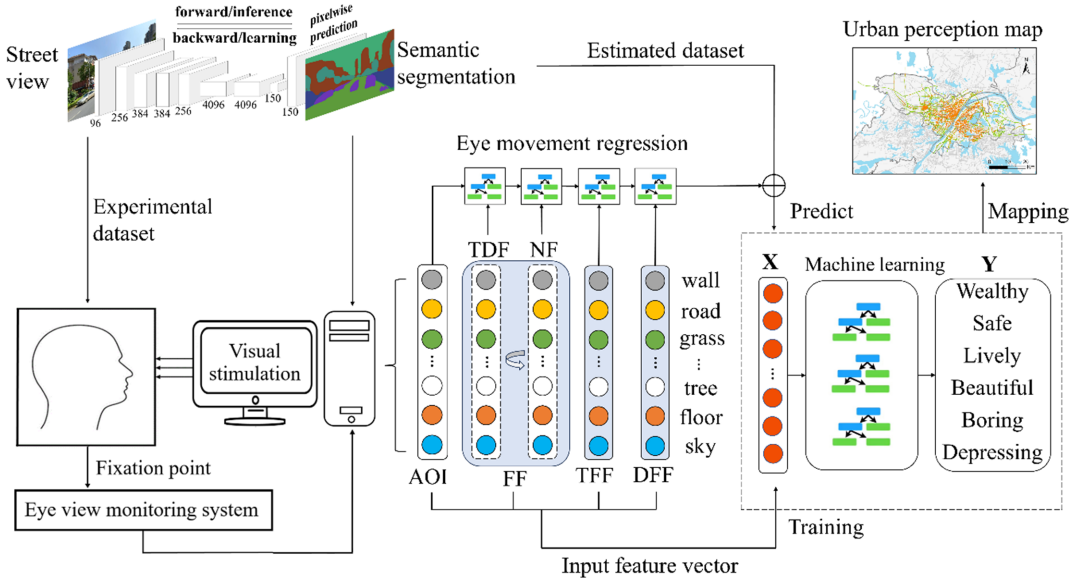

FIGURE 8 Construction process of urban perception model.

5.1 | Training on the experimental dataset

This study utilizes semantic segmentation results of street view images as input for eye movement data, defining the AOI for eye movement stimuli. Three eye movement metrics, including, TFF, DFF, and FF, are extracted as

eye movement feature data. Additionally, the area shares of features obtained through semantic segmentation of street view images serve as semantic feature data, contributing to the construction of a feature dataset that influences urban perception scores. Finally, we used the random forest algorithm to regress the predicted perception scores by randomly partitioning the dataset, of which 80% was divided into the training set and the remaining 20% of the test set was used as out-of-bag (OOB) data to evaluate the model error. To evaluate the impact of eye movement feature data on the prediction accuracy and stability of the perceptual model, this study employs the random forest algorithm to train two separate datasets. One dataset consists solely of semantic features, while the other includes only eye movement features. Comparative experiments are conducted to assess the disparities between models built on the two separate datasets and the models that integrate both eye movement and semantic features. In the random forest model for the prediction of perceptual scores, this study uses Pearson correlation coefficient (γ), root mean square error (RMSE), mean absolute error (MAE) and standard R^2 to assess the fitting accuracy of the model.

Table 4 shows the accuracy of predicting six city perceptions based on three datasets of OOB: solely semantic feature data, solely eye movement feature data, and both eye movement and semantic features data. The random forest model incorporating both eye movement and semantic feature data demonstrates higher accuracy in predicting city perceptions compared with using semantic or eye movement features alone. In the test set, the Pearson's coefficient exceeds 60%, while the absolute average error between predicted and actual values is limited to within 11. Furthermore, the R^2 surpasses 0.53 and reaches a maximum of 0.62 for lively perception prediction.

Models that incorporate eye movement features, either alone or in combination with semantic features, exhibit a notable distinction in terms of data sources when compared with previous studies focused on single-factor, homogeneous contexts, and coarse-grained urban perception (Wu et al., 2022). While the increase in accuracy is only marginal, the utilization of eye movement data introduces a novel dimension to the research. The premise underlying studies that solely rely on semantic segmentation assumes that human gaze is evenly distributed across all elements of a given feature. However, this assumption does not align with the principles of human perception. Our eye movement studies have demonstrated that human gaze on street view images is not uniformly distributed. Instead, individuals tend to focus their attention on specific areas before forming evaluations. Therefore, we contend that incorporating eye movement data into related studies is a more scientifically sound approach. By considering the actual focal points of participants' gaze, we can gain deeper insights into the evaluation process and improve the accuracy of our analyses.

5.2 | Regression on the estimated dataset

Due to the constraints imposed by the eye tracking device, there were limitations on the number of photos and subjects that could be included in the eye movement experiment. To address this constraint, we need to use semantic segmentation features as predictors for the eye movement data on the predicted dataset. By applying the random forest algorithm to analyze patterns and trends in eye movement data and semantic segmentation data, we can derive insights into the visual elements and regions that capture attention in a particular street scene image. As a result, we can effectively regress the eye movement features associated in the estimated dataset, pertaining to street images. We divided the experimental eye movement dataset in an 8:2 ratio, of which 80% was used to train the random forest and the remaining 20% was used as test data for model validation. Model fitting results were quantified using RMSE, MAE and standard R^2, to quantify the accuracy between predicted and true values.

To minimize potential errors, we utilize the predicted values of TDF and the NF in order to calculate the predicted values of FF using formula 2. The results of fitting the four eye movement metrics based on the random forest model for tTDF, NF, TFF, and DFF are shown in Table 5. It can be seen that the MAE between the predicted and true values of the four eye movement metrics models is within 0.01 (1%), the RMSE is within 0.05 (5%), and $\sum_{i=1}^{n} \hat{y}_i = 1$, and R^2 are all above 0.65. This indicates that the accuracy of fitting human eye movement

TABLE 4 Comparison of fitting accuracy of random forest models based on three different data sets of OBB.

	Only input semantic features				Only input eye movement features				Both eye movements and semantic features			
Perceptions	R^2	RMSE	MAE	γ	R^2	RMSE	MAE	γ	R^2	RMSE	MAE	γ
Wealthy	0.5525	15.91	13.07	53.43%	0.5485	14.83	11.61	67.59%	0.6102	12.44	10.45	75.83%
Safety	0.4421	12.04	10.19	57.80%	0.4973	12.5	10.04	59.34%	0.5361	10.19	9.14	69.81%
Lively	0.5827	16.15	13.54	52.60%	0.5651	16.19	12.42	70.38%	0.6243	12.25	10.66	78.62%
Beautiful	0.4743	15.64	12.26	51.16%	0.5359	15.52	11.38	60.70%	0.5646	13.09	10.8	76.03%
Boring	0.5017	12.83	10.65	59.35%	0.5121	12.39	10.45	56.15%	0.5544	10.75	9.18	61.66%
Depressing	0.5591	13.32	11.46	58.12%	0.5863	13.16	10.28	65.56%	0.6165	10.84	9.34	68.36%

TABLE 5 Results of fitting the four eye movement metrics based on the random forest model.

	RMSE		MAE		R^2	
	Train set	Test set	Train set	Test set	Train set	Test set
TDF	0.0146	0.0395	0.0026	0.0070	0.8497	0.6879
NF	0.0145	0.0327	0.0027	0.0072	0.8258	0.6720
TFF	0.0168	0.0444	0.0033	0.0093	0.8035	0.6562
DFF	0.0110	0.0290	0.0024	0.0065	0.8580	0.6971

data based on the percentage of different features in each street view image in the study area is very high and can represent the eye movement characteristics on different visual elements of the street view images.

5.3 | Prediction results and cartography of urban perception

Figure 9 illustrates the sentiment map of Wuhan city across six dimensions, generated using the random forest model. The sentiment depth of each unit is calculated by averaging the values obtained from the four directions of streetscape sampling points distributed along the streets. The map reveals that within the Third Ring Road, Wuhan city exhibits higher levels of wealth, safe, and liveliness compared with areas outside the ring. This is particularly evident in the densely developed urban regions within the Third Ring Road, where the perception of 'safe' is notably elevated. Wuchang, Jianghan, and Jiang'an are the oldest urban areas of Wuhan, with traditionally dense residential areas and more mature commercial and business districts, so the perceived levels of wealthy, safe, and lively are higher than in the surrounding urban areas. Guanggu, as an emerging urban area in Wuhan, has a rapidly growing economy and population, and the urban area is developing rapidly, so it also scores higher on the above three perceptions. Regarding the perception of beauty, the outlying areas of Hongshan, as well as Dongxihu, which are situated at a considerable distance from the city center, exhibited higher scores. However, certain scattered streets in Wuchang, Jianghan, and Jiang'an, representing the older urban areas, display lower levels of beauty perception. These lower levels can be attributed to factors, such as population density, traffic congestion, overloaded infrastructure, and the outdated condition of buildings in need of renovation. The high level of boring perception is mainly concentrated in areas, such as arterial roads, elevated roads, and tunnels away from residential areas, where the building structure is relatively homogeneous and urban functions are simple. Nevertheless, the study area as a whole is characterized by a low overall perception of depressing, primarily due to the presence of wide-view arterial roads at the sampling points. It is also related to the quality of the street view images, which were mostly taken in good lighting and weather conditions, which may have had a negative impact on the perception of depressing.

Figure 10 shows the tag map based on the main perceptual distributions within the study area of Wuhan city. We employed the generalized kriging interpolation method to interpolate the line distributions of the six perceptions into a map format. Through superposition analysis, we identified and retained the dimensions with the highest ratings from the six perceptions. This approach allowed us to capture the spatial distribution of the most representative perceptions in each region and present them in the form of tag map. This visually appealing tag map serve to enhance accessibility and engagement, enabling readers to easily grasp the predominant perceptions within each area. The analysis reveals that the tags indicating "wealthy", "safe", and "lively" are predominantly concentrated within the primary urban areas encompassed by the Third Ring Road. In contrast, the tags denoting "beautiful" are mainly distributed near locations such as the East Lake Scenic Area in Wuchang, South Lake Avenue in Hongshan, and Jinyin Lake Scenic Area in Dongxihu. On the other hand, the tags associated with "boring" and "depressing" are less widespread, primarily concentrated in the suburban regions distant from the

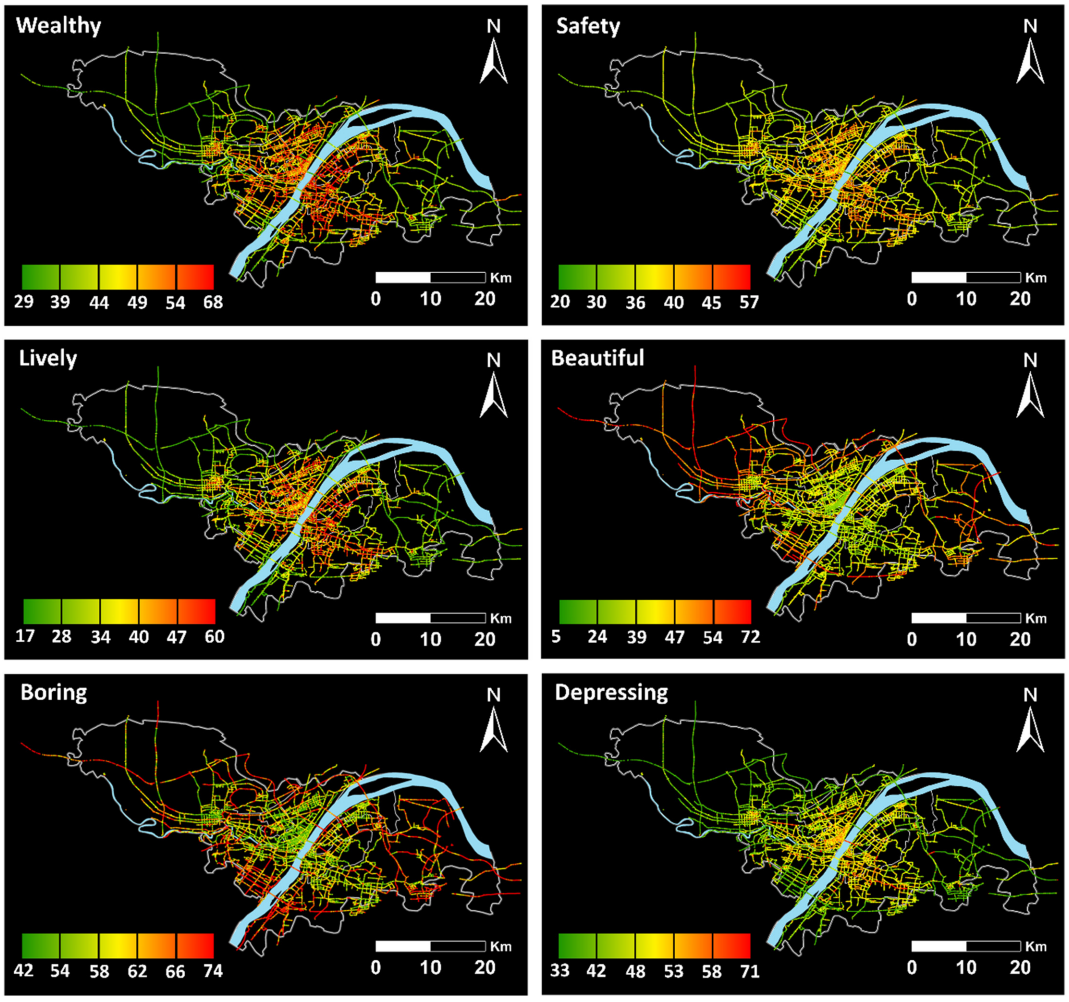

FIGURE 9 Fusing eye movement feature data and semantic feature data to fit six dimensions of urban perception results from 8 main urban areas of Wuhan.

city center and undeveloped areas. The distribution pattern of these tags aligns closely with the observations depicted in Figure 9.

6 | DISCUSSION

The perception of urban environments primarily relies on human subjects, with vision serving as the primary sensory channel through which individuals perceive the city. Previous studies on urban perception have neglected human visual attention, and the research in this article makes up for this deficiency. Nevertheless, the following aspects deserve further investigation:

1. The selection of eye movement metrics. In this article, TFF, DFF, and FF derived from TDF and NF are selected. However, it is important to note that there are numerous other eye movement metrics available, including fixation density, regression time, pupil diameter, and many more. Are there also correlations between these eye movement metrics and urban perception? Can the accuracy of urban perception be

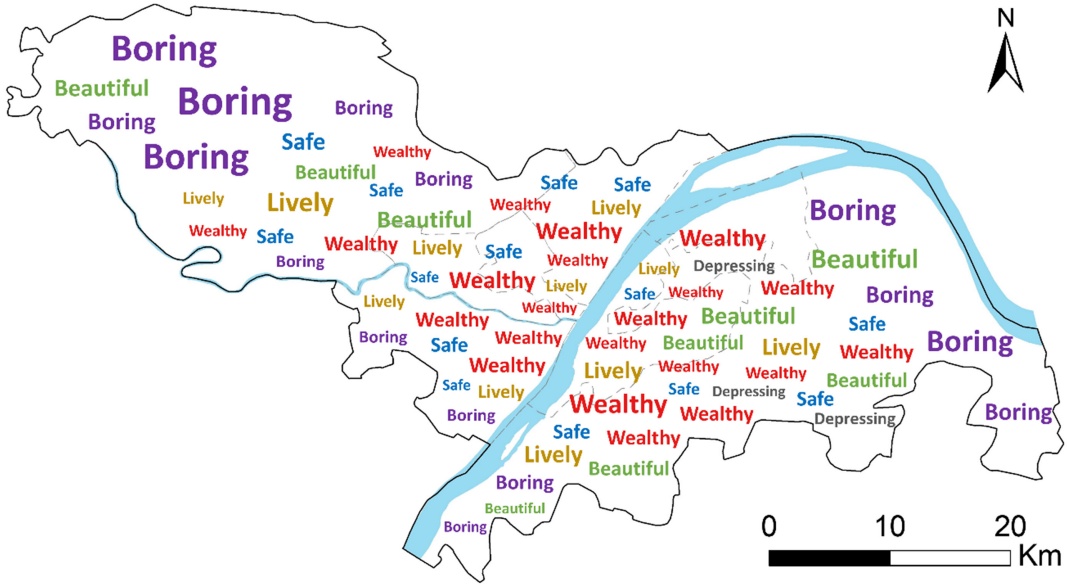

FIGURE 10 A tag map with integrated distribution of six perceptual dimensions.

improved by incorporating more eye movement metrics? These questions deserve further investigation. Furthermore, this study solely employs random forest model to learn the semantic features for predicting eye movement data on the designated dataset. Future research endeavors could explore the inclusion of additional variables that potentially influence eye movement metrics, such as color information, spatial depth perception, and other relevant factors.

2. Driving factors of urban perception differences. This article only analyzes the correlation between the eye movement indices of different visual elements and urban perception. The driving factors behind the urban perception differences need to be further investigated. Indeed, Each person has unique patterns and behaviors when it comes to eye movements, influenced by factors such as cognitive processes, attentional focus, and personal experiences (Wu et al., 2022). Recognizing and accounting for this variability is crucial for exploring human urban perceptual biases from eye movement studies in subsequent studies.
3. Selection of perceptual dimensions. In this article, we refer to the MIT Media Lab's "Place Pulse 2.0" project and select six perceptions: wealthy, safe, lively, beautiful, boring, and depressing (Dubey et al., 2016). Do these perceptions cover all the dimensions of people's perceptions of cities? Does the choice of perceptual dimensions of cities need to be differentiated in different countries and regions? These questions should also be discussed.
4. Limitations of street view images. Street view images are static, and they are often taken from angles with better light vision, while human perception of the city should be all-round, such as real-time traffic, cultural heritage, climate, and other factors beyond static images. Future research could even employ outdoor eye tracking devices to collect urban perception and eye tracking data from real street environments.

7 | CONCLUSIONS

The contributions of this article include:

1. Drawing upon the data-driven experimental paradigm, this study extracts eye movement metrics using eye tracking technology and image semantic segmentation. It integrates and correlates these features

with pertinent theories from psychology to analyze the interactive effects of various visual elements on urban perception. By considering cognitive patterns reflected in eye movements, we gain insights into the relationship between visual elements and perception, determining how they impact fixation and correlate with perception scores. Meanwhile, this analysis identifies which physical environments negatively or positively influenced place perception.

2. This study proposes a more scientific and reliable method of measuring urban perception by incorporating eye movement data. The perceptual model integrates semantic and eye movement features to provide a comprehensive understanding of the unconscious visual and cognitive behaviors of individuals that are difficult to express verbally when perceiving the physical environment. The model integrates human visual attention and psychological motivation by recording participants' visual attention and extracting three eye movement metrics for various visual elements. Consequently, it advances the study of environmental perception and reveals the complex dynamics of how humans perceive and interact with their surroundings.
3. Based on the above studies, it is foreseeable that with breakthroughs in eye tracking technology and behavioral cognition theory, there is a promising opportunity to leverage spatial information data, individual physiological, and psychological data for assessing and quantifying urban perception. This will enable personalized urban planning and design, supported by the theory of the influence of individual psychological factors on urban perception, ultimately improving the quality of life and experience for residents.

ACKNOWLEDGMENTS

The authors would take to thank editors and anonymous reviewers for the useful comments on the manuscript.

CONFLICT OF INTEREST STATEMENT

No potential conflict of interest was reported by the author(s).

DATA AVAILABILITY STATEMENT

The data and codes that support the findings of this study are openly available in figshare at https://figshare.com/s/23fdb9d87f09c5a6bcf6.

ORCID

Nai Yang https://orcid.org/0000-0001-5306-1163
Zhiwei Wei https://orcid.org/0000-0002-3494-3686

REFERENCES

Cavicchio, F., Melcher, D., & Poesio, M. (2014). The effect of linguistic and visual salience in visual world studies. *Frontiers in Psychology*, *5*, 176. https://doi.org/10.3389/fpsyg.2014.00176

Chen, Y., Jia, B., Jing, W., Liu, X., & Luo, T. (2022). Temporal and spatial attractiveness characteristics of Wuhan urban Riverside from the perspective of traveling. *Land*, *11*, 1434. https://doi.org/10.3390/land11091434

Dadvand, P., Bartoll, X., Basagaña, X., Dalmau-Bueno, A., Martinez, D., Ambros, A., Cirach, M., Triguero-Mas, M., Gascon, M., Borrell, C., & Nieuwenhuijsen, M. J. (2016). Green spaces and general health: Roles of mental health status, social support, and physical activity. *Environment International*, *91*, 161–167. https://doi.org/10.1016/j.envint.2016.02.029

Dong, W., Liao, H., Zhan, Z., Liu, B., Wang, S., & Yang, T. (2019). New research progress of eye tracking-based map cognition in cartography since 2008. *Acta Geographica Sinica*, *74*, 599–614. https://doi.org/10.11821/dlxb201903015

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I, 14, 196–212. https://doi.org/10.1007/978-3-319-46448-0_12

Fabrikant, S. I., Hespanha, S. R., & Hegarty, M. (2010). Cognitively inspired and perceptually salient graphic displays for efficient spatial inference making. *Annals of the Association of American Geographers*, *100*, 13–29. https://doi.org/10.1080/00045600903362378

Fairbairn, D., & Hepburn, J. (2023). Eye-tracking in map use, map user and map usability research: What are we looking for? *International Journal of Cartography*, *9*, 231–254. https://doi.org/10.1080/23729333.2023.2189064

Garza, R., Heredia, R. R., & Cieslicka, A. B. (2016). Male and female perception of physical attractiveness. *Evolutionary Psychology*, *14*, 1–16. https://doi.org/10.1177/1474704916631614

Gombrich, E. H. (1995). *The story of art*. Phaidon.

Greene, M. R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*, 464–472. https://doi.org/10.1111/j.1467-9280.2009.02316.x

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

Guo, S., Zhao, N., Zhang, J., Xue, T., Liu, P., Xu, S., & Xu, D. (2017). Landscape visual quality assessment based on eye movement: College student eye-tracking experiments on tourism landscape pictures. *Resources Science*, *39*, 1137–1147. https://doi.org/10.18402/resci.2017.06.13

Helbich, M., Yao, Y., Liu, Y., Zhang, J., Liu, P., & Wang, R. (2019). Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environment International*, *126*, 107–117. https://doi.org/10.1016/j.envint.2019.02.013

Herzog, T. R. (1992). A cognitive analysis of preference for urban spaces. *Journal of Environmental Psychology*, *12*, 237–248. https://doi.org/10.1016/S0272-4944(05)80138-0

Kelling, G. L., & Coles, C. M. (1997). *Fixing broken windows: Restoring order and reducing crime in our communities*. Simon and Schuster.

Kiefer, P., Giannopoulos, I., & Raubal, M. (2013). *Using eye movements to recognize activities on cartographic maps. Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. https://doi.org/10.1145/2525314.2525467

Kiefer, P., Giannopoulos, I., Raubal, M., & Duchowski, A. (2017). Eye tracking for spatial research: Cognition, computation, challenges. *Spatial Cognition & Computation*, *17*, 1–19. https://doi.org/10.1080/13875868.2016.1254634

Klaib, A. F., Alsrehin, N. O., Wasen, Y. M., Bashtawi, H. O., & Magableh, A. A. (2021). Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and internet of things technologies. *Expert Systems with Applications*, *166*, 114037. https://doi.org/10.1016/j.eswa.2020.114037

Krassanakis, V., & Cybulski, P. (2019). A review on eye movement analysis in map reading process: The status of the last decade. *Geodesy and Cartography*, *68*, 191–209. https://doi.org/10.24425/gac.2019.126088

Liao, H., Dong, W., Huang, H., Gartner, G., & Liu, H. (2018). Inferring user tasks in pedestrian navigation from eye movement data in real-world environments. *International Journal of Geographical Information Science*, *33*, 739–763. https://doi.org/10.1080/13658816.2018.1482554

Lin, N., & Liu, F. (2012). Analysis of Wuhan urban characteristics and urban spatial planning. *Applied Mechanics and Materials*, *174*, 2516–2519. https://doi.org/10.4028/www.scientific.net/AMM.174-177.2516

Lindal, P. J., & Hartig, T. (2013). Architectural variation, building height, and the restorative quality of urban residential streetscapes. *Journal of Environmental Psychology*, *33*, 26–36. https://doi.org/10.1016/j.jenvp.2012.09.003

Noland, R. B., Weiner, M. D., Gao, D., Cook, M. P., & Nelessen, A. (2016). Eye-tracking technology, visual preference surveys, and urban design: Preliminary evidence of an effective methodology. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, *10*, 98–110. https://doi.org/10.1080/17549175.2016.1187197

Ohm, C., Müller, M., & Ludwig, B. (2016). Evaluating indoor pedestrian navigation interfaces using mobile eye tracking. *Spatial Cognition & Computation*, *17*, 89–120. https://doi.org/10.1080/13875868.2016.1219913

Oki, T., & Kizawa, S. (2021). Evaluating visual impressions based on gaze analysis and deep learning: A case study of attractiveness evaluation of streets in densely built-up wooden residential area. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 887–894. https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-887-2021

Pieters, R., & Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of Marketing*, *68*, 36–50. https://doi.org/10.1509/jmkg.68.2.36.27794

Porzi, L., Bulò, S. R., Lepri, B., & Ricci, E. (2015). Predicting and understanding urban perception with convolutional neural networks. *Proceedings of the 23rd ACM International Conference on Multimedia*, 139–148. https://doi.org/10.1145/2733373.2806273

Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PLoS One*, *8*, e68400. https://doi.org/10.1371/journal.pone.0068400

Shao, Z. (2019). *Cognitive psychology: Theories, experiments and applications (Third Edition)*. Shanghai Education Publishing House.

Tamir, M., & Robinson, M. D. (2007). The happy spotlight: Positive mood and selective attention to rewarding information. *Personality and Social Psychology Bulletin*, *33*, 1124–1136. https://doi.org/10.1177/0146167207301030

Wang, L., Han, X., He, J., & Jung, T. (2022). Measuring residents' perceptions of city streets to inform better street planning through deep learning and space syntax. *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 215–230. https://doi.org/10.1016/j.isprsjprs.2022.06.011

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

Wu, Y., Dong, W., & Zhang, W. (2022). Research on influencing factors of built environment perception in neighborhoods: Evidence from behavioral experiment. *City Planning Review*, *46*, 99–109. https://doi.org/10.11819/cpr20221206a

Yang, N., Guojia, W., MacEachren, A. M., Pang, X., & Fang, H. (2023). Comparison of font size and background color strategies for tag weights on tag maps. *Cartography and Geographic Information Science*, *50*, 162–177. https://doi.org/10.1080/15230406.2022.2152098

Yao, Y., Liang, Z., Yuan, Z., Liu, P., Bie, Y., Zhang, J., Wang, R., Wang, J., & Guan, Q. (2019). A human-machine adversarial scoring framework for urban perception assessment using street-view images. *International Journal of Geographical Information Science*, *33*, 2363–2384. https://doi.org/10.1080/13658816.2019.1643024

Zemblys, R., Niehorster, D. C., Komogortsev, O., & Holmqvist, K. (2018). Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, *50*, 160–181. https://doi.org/10.3758/s13428-017-0860-3

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, *180*, 148–160. https://doi.org/10.1016/j.landurbplan.2018.08.020

Zhang, N., Zheng, X., & Wang, X. (2022). Assessment of aesthetic quality of urban landscapes by integrating objective and subjective factors: A case study for riparian landscapes. *Frontiers in Ecology and Evolution*, *9*, 735905. https://doi.org/10.3389/fevo.2021.735905

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2881–2890. https://doi.org/10.48550/arXiv.1612.01105

Zheng, S. (2020). *Personalized map cognition and eye movement analysis method*. Publishing House of Electronics Industry.

How to cite this article: Yang, N., Deng, Z., Hu, F., Chao, Yi, Wan, L., Guan, Q., … Wei, Z. (2024). Urban perception by using eye movement data on street view images. *Transactions in GIS*, *00*, 1–22. https://doi.org/10.1111/tgis.13172