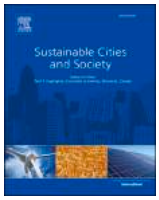
Sustainable Cities and Society 86 (2022) 104140

Contents lists available at ScienceDirect

Sustainable Cities and Society

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



Measuring visual walkability perception using panoramic street view images, virtual reality, and deep learning

Yunqin Li, Nobuyoshi Yabuki *, Tomohiro Fukuda

Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University, Japan

ARTICLE INFO

Keywords:
Visual walkability perception (VWP)
Panoramic street view images
Virtual reality
Deep learning
Built environment

ABSTRACT

Measuring perceptions of visual walkability in urban streets and exploring the associations between the visual features of the street built environment that make walking attractive to humans are both theoretically and practically important. Previous studies have used either environmental audits and subjective evaluations that have limitations in terms of cost, time, and measurement scale, or computer-aided audits based on natural street view images (SVIs) but with gaps in real perception. In this study, a virtual reality panoramic image-based deep learning framework is proposed for measuring visual walkability perception (VWP) and then quantifying and visualizing the contributing visual features. A VWP classification deep multitask learning (VWPCL) model was first developed and trained on human ratings of panoramic SVIs in virtual reality to predict VWP in six categories. Second, a regression model was used to determine the degree of correlation of various objects with one of the six VWP categories based on semantic segmentation. Furthermore, an interpretable deep learning model was used to assist in identifying and visualizing elements that contribute to VWP. The experiment validated the accuracy of the VWPCL model for predicting VWP. The results represent a further step in understanding the interplay of VWP and street-level semantics and features.

1. Introduction

Walkable streets and urban areas have been shown to promote social and economic prosperity in communities (Duncanet al., 2011). The walkability of streets has become one of the key factors that urban planners and designers consider when designing and building urban communities (Zhou et al., 2019). The term “walkability” is commonly defined as the extent to which a built environment is friendly and inviting to people (Abley & Hill, 2005) who walk out of necessity, by choice, or for social purposes (Cerin et al., 2007) and increases the willingness of pedestrians to walk in a street space (Ewing & Handy, 2009). Individual differences are inevitable in the decision to walk and choice of walking behavior in urban communities, so walkability relies heavily on human perception (Wang & Yang, 2019). Walkability is a subjective concept that describes the quality of the environment and is influenced by individual perceptions of the environment as a place suitable for walking, and thus is difficult to measure objectively and quantitatively (Wang & Yang, 2019). However, objectively quantifying walkability at the street level and gaining knowledge about how the physical setting and visual information of the built environment affect the pedestrian experience can aid in the development of design strategies for walkable neighborhoods (Arellana et al., 2020).

With the rapid development of mapping services, the large-scale quantitative measurement of walkability from street view images (SVIs) based on computer vision technology has become a trend in the era of big data (Arellana et al., 2020; Blečić et al., 2018; Fan et al., 2018; Ki & Lee, 2021; Li et al., 2020; Wang & Yang, 2019; Zhou et al., 2019). Because visual information is the cornerstone of human perception of the environment, a visual walkability and visual walkability index based on SVIs has been proposed as a measure of pedestrian psychological and subjective comfort, which can be approximated from psychological greenery, visual crowdedness, outdoor enclosures, and visual pavement (Zhou et al., 2019). However, calculations that quantify the physical characteristics of these streetscape images are hardly representative of subjective on-site perceptions for visual walkability (Li et al., 2020). Therefore, methods are needed for measuring visual walkability perception (VWP) and exploring its relationship with visual elements of the street built environment.

In early studies, one approach taken by many urban designers and researchers was to conduct interviews and observe walking activities to

* Corresponding author.
E-mail address: yabuki@see.eng.osaka-u.ac.jp (N. Yabuki).

https://doi.org/10.1016/j.scs.2022.104140
Received 11 May 2022; Received in revised form 11 July 2022; Accepted 18 August 2022
Available online 21 August 2022

find connections between physical appearance and walking behavior based on logical reasoning, thus measuring walkability subjectively rather than quantitatively (Saadi et al., 2022; Wang & Yang, 2019). In another approach, many indices for evaluating walkability have been created, each containing some terminology related to walkability in tools or checklists to assess walkability in pedestrian environments through field surveys and questionnaires (Campisi et al., 2021; Wang & Yang, 2019). These environmental audits and subjective evaluation methods have obvious advantages in terms of approximating reality but have limitations in terms of cost, time, and measurement scale. Further, their validity is difficult to verify in controlled experiments.

In recent years, computer-aided auditing methods based on natural SVIs have emerged and are capable of objectively auditing a large number of walkability-related streetscape elements (Blečić et al., 2018), mitigating the shortcomings of previous manual evaluations with insufficient sample size and making it easy to apply the methods for comparisons across cities (Li et al., 2020). However, these methods have gaps in capturing and visually revealing real on-site walkability perceptions (Kim & Lee, 2022). For example, some perceptions that require a sense of space and similar levels of realism such as human scale, enclosures are hard to be experienced on browser-based auditing of natural SVIs. Meanwhile, due to the limitation of natural SVI camera angles, it is difficult to avoid different perceptual evaluation results corresponding to multiple natural SVIs with different perspectives at the same location. In addition, due to the opaque working mechanism of deep neural network models, ordinary urban planners and designers do not have access to the computer's internal process for making decisions, increasing the difficulty of building trust and promoting the computer model's output, especially when faced with tasks involving subjective perceptions such as visual walkability that are inherently difficult to interpret and verify (Guidotti et al., 2018; Ibrahim et al., 2019).

Virtual reality (VR) technologies and deep convolutional neural network (DCNN)-based deep learning algorithms overcome the limitations of previous approaches and show great potential in visual walkability perception studies. Immersion in a street environment consisting of 360-degree SVIs through a head-mounted VR device is more cost effective and has feasibility and effectiveness close to that of on-site perception audits (Kim & Lee, 2022; Lee & Kim, 2021). DCNN-based image classification can measure, classify, and predict VWP based on the deep features of panoramic SVIs on a large scale (Hu et al., 2020), and a semantic segmentation algorithm can be used for batch analysis of the elemental share of SVIs (Yin & Wang, 2016), which is necessary for quantifying the relationship between the built environment and VWP (Li et al., 2022). Interpretable artificial intelligence provides an interface between designers and computer-aided decision models for VWP metrics, which helps us to understand, learn, and validate the decision process of VWP measurement models (Min et al., 2020).

In the present study, a data-driven deep learning framework is constructed for VR-based VWP measurement with panoramic SVIs. Our research objectives are to (1) propose a VR-based pairwise comparison approach for VWP scoring in six categories, namely, walkable, feasibility, accessibility, safety, comfort, and pleasurability; (2) design a VWP classification deep multitask learning (VWPCL) model with a tailored dataset; (3) use a stepwise multiple linear-regression model to analyze the relationship between object ratios and VWP scores obtained by semantic segmentation of panoramic SVIs from a macroscale perspective; (4) visualize the objects contributing to the VWP evaluation using interpretable deep learning from a microscale perspective; and (5) validate the effectiveness of our VWPCL model and its interpretable deep learning results.

Our main contributions described in this paper can be summarized as follows:

- A VR panoramic-based deep learning framework for VWP measurement was proposed. This is a new paradigm for observing, perceiving, auditing, and understanding the street built environment features and subjective visual perceptions based on the big data of panoramic SVIs. Compared to browser-based evaluation, immersive VR visualization helped raters make evaluations that were close to their real on-site perceptions. In addition, the VR panorama-based audit also solved the scoring bias of the natural SVIs based on different views of the same location with more consistent results.
- A VWPCL model based on DCNN was trained to classify and predict VMP evaluation results and verified with on-site auditing results. Furthermore, a tailored image dataset for training VMP classifiers was built.
- The association between the street visual elements and each specific visual walkability perception was explored using multiple linear regression. In addition, we applied a Grad-CAM model to visualize the saliency areas of each VWP category and evaluated the consistency of the contributing visual elements identified in the questionnaire with the activation heat map of the deep learning model. Overall, these two methods provide both macro- and micro-scale perspectives for observing and interpreting the relationship between VWP and the elements of a street built environment.

The rest of this manuscript is organized as follows. Section 2 reviews related work and Section 3 explains the framework and steps of our research method. Section 4 describes the experiment and the results of our proposed methods and method verification. Section 5 summarizes our findings and contributions and discusses the limitations and future work, and Section 6 provides concluding remarks.

2. Related work

2.1. Visual walkability perception

Over the past several decades, walkability has been of research interest in the fields of urban design, transport, and public health. Walkability describes the extent to which a street's environment and form support and encourage walking (Southworth, 2005). Walkability is a concept involving subjective perception that is difficult to quantify, but there have been many attempts to evaluate walkability quantitatively from multiple aspects (Aghaabbasi et al., 2018; Chan et al., 2021; Ortega et al., 2021; Yameqani & Alesheikh, 2019), though no consensus has yet been reached. Li et al. (2020) divided street walkability into physical walkability and perceived walkability and investigated the extent to which physical walkability is representative of perceived walkability. Alfonzo (2005) proposed a hierarchy of walking needs as a subjective walkability evaluation index, which represents the precursors to the process of deciding to walk. Zhou et al. (2019) first introduced the concept of visual walkability based on SVIs and proposed a visual walkability index as a quantitative indicator for visual walkability evaluation. The overall walking behavior generally evokes a combination of visual and non-visual perceptions, influenced by the built environment, weather, and human behavior. In the present paper, VWP refers to whether the built environment of a neighborhood visually encourages people to walk, with an emphasis on the influence of visual elements on walking intentions.

Inspired by Alfonzo (2005), VWP consists of one visual walkable indicator and a hierarchy of five visual indicators of walking needs, namely, feasibility, accessibility, safety, comfort, and pleasurability. In other words, the street-level VWP has six categories. The visual walkable indicator describes the overall visual impression of whether a location is supportive of walking. Feasibility indicates the incentive factors for the occurrence of walking, which can be affected by types of land use and diversity of facilities. Accessibility refers to visible barriers to walking, such as dead-end streets and restricted access. Safety describes whether a street space appears safe from crime and traffic accidents. This is related to visual factors such as graffiti, litter, and abandoned buildings. Comfort, which describes whether the street space improves the pedestrian walking experience, is associated with street furniture,

sidewalk width, urban design amenities, and barrier-free facilities. Pleasurability measures the level of appeal, typically meaning the extent to which public spaces are diverse, lively, enjoyable, and interesting to walk in. Fig. 1 shows how the key visual factors were screened for measuring VWP in the hierarchy of walking needs.

2.2. Panoramic SVIs and virtual-reality-based visual perception

The extensive availability of SVIs has become one of the primary data sources in urban research to study how the physical and objective attributes of streets affect peoples' subjective perceptions (Zhang et al., 2021). Fully open-access SVI data provide high-resolution, real-world, street-level image data for visual perception studies of urban street built environments (He & Li, 2021). Construction of an SVI database is cost effective with a straightforward workflow. Several studies have attempted to establish a link between SVI and peoples' subjective perceptions using street view scoring (Li et al., 2015), using such indicators of perceptual assessment as beautiful, lively, safe, wealthy, boring, and depressing (Salesses et al., 2013; Zhang et al., 2018) and beautiful, quiet, and pleasant (Quercia et al., 2014). The natural SVIs used in these studies provide an approximate view of pedestrians, but they have difficulty in demonstrating a sense of reality and space in the actual built environment of the street (Kim & Lee, 2022).

VR technology with panoramic SVIs provides a virtual world of street built environments where pedestrians can be immersed in, present in, and interactive with the street environment (Bellazzi et al., 2022). High-resolution panoramic SVIs from a street view service have consistent and controlled quality in image processing for VR preparation. The mobile-based head-mounted display (HMD) with motion sensors allows users to have a 360-degree view of the virtual world and greatly increases the realism and presence of the VR street view environment. This helps participants to meticulously capture multi-level information of the street view environment and make subjective perceptual judgments that closely reflect the real environment. A VR-based cityscape visual analytics system has been built to collect and analyze perceptual data on visual attention (Zhang & Zhang, 2021). Some scholars have attempted to explore residents' perceptions of the appearance of urban construction or to audit the streetscape by displaying 360-degree videos of real street environments using mobile VR platforms (Kim & Lee, 2022; Mouratidis & Hassan, 2020). Inspired by these works, we developed a workflow for measuring and evaluating VWP based on immersive VR and panoramic SVI.

2.3. DCNN-based deep learning methods using SVIs to measure and interpret VWP

Recent advances in DCNN-based deep learning technology on SVIs, including image classification, semantic segmentation, and interpretable methods, have opened up new possibilities for measurement and interpretation of VWP.

The use of traditional machine learning methods, such as support vector machines (SVMs), in image classification started long ago (Zhang et al., 2018). However, these methods suffer from disadvantages such as feature extraction and selection being time-consuming and performance being disconnected from practical standards (Rawat & Wang, 2017). In recent years, DCNNs, such as VGG-16 (Simonyan & Zisserman, 2014), ResNet (He et al., 2016), and the densely connected convolutional network (DenseNet) (Huang et al., 2017), have been widely used for evolving image classification tasks and have achieved remarkable performance. Several studies used image classification methods to evaluate and predict human perception of a street. Zhang et al. (2018) trained an SVM classifier on the Place Pulse 2.0 dataset (Dubey et al., 2016) to predict human perceptions of six indicators, which was followed by Min et al. (2020), who proposed a ranking SVM-based model to learn visual urban perception. Verma et al., (2020)0 used an image classifier model pre-trained on the Places365 dataset to extract high-level features of a street view to map human visual perception.

In addition, a large number of studies have been conducted to audit the street built environment using DCNN-based semantic segmentation of SVIs and interpret the association between street-level physical components and human perceptions (Ma et al., 2021; Quercia et al., 2014; Tang & Long, 2019; Yao et al., 2019; Zhang et al., 2018). In these studies, semantic segmentation is usually first used to extract the physical components of the street from millions of SVIs, and then urban

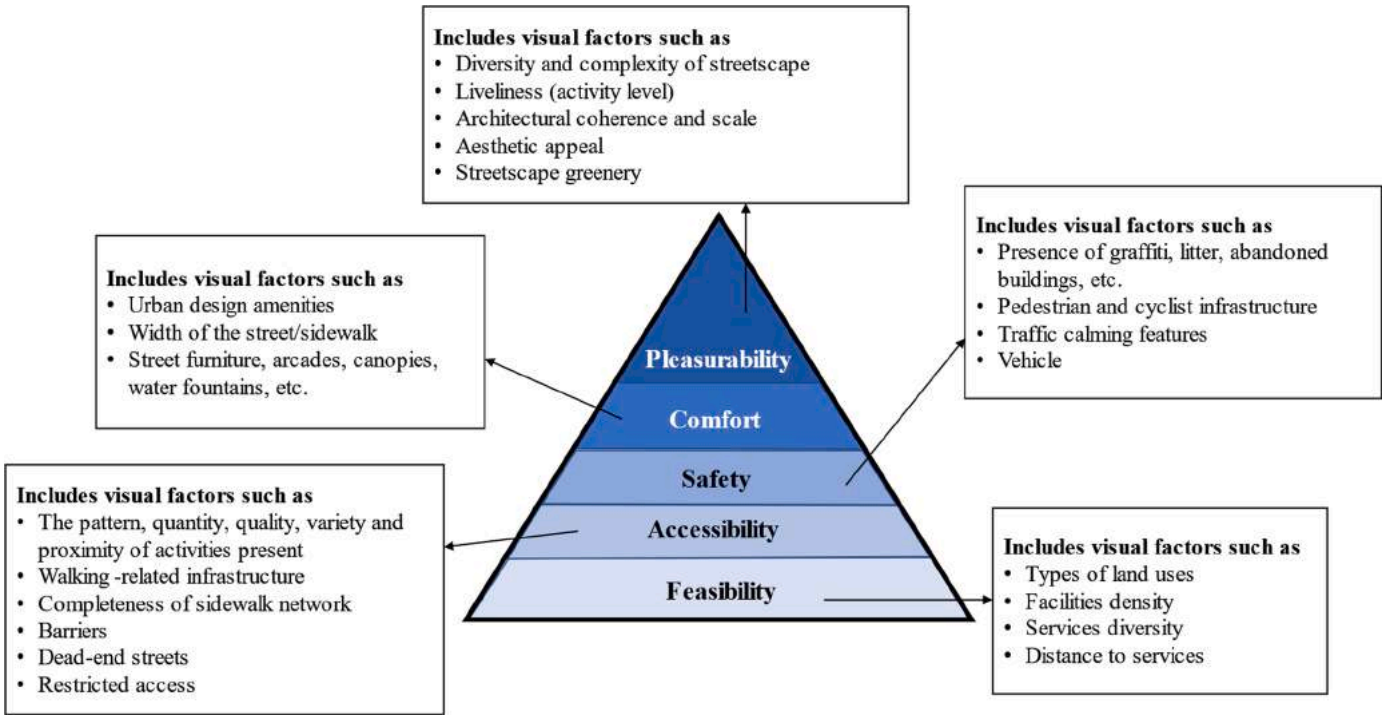


Fig. 1. Visual factors in the hierarchy of walking needs.

perception analysis is implemented through a scoring system. DeepLab-v3+ (Chen et al., 2018) is one of the high-accuracy semantic segmentation model that uses atrous convolution layers and a simple decoder module to provide fine-grained segmentation (Li et al., 2022).

Although DCNN-based approaches have shown powerful feature learning capabilities and achieved state-of-the-art performance in computer vision tasks such as image classification, the interpretability of DCNNs has often been criticized by stakeholders (Linardatos et al., 2020), especially for applications such as measurement for complex perception. Thus, there is an urgent need for stakeholders to find a way to understand and explain how DCNNs learn, why classification DCNNs predict what they predict, and whether the decisions they make are reliable (Fu et al., 2020). This is because these image classification networks often appear to be complex black boxes with a large number of uninterpretable parameters (Chen et al., 2020).

For models with complex results, such as DCNNs, many classical post hoc interpretation methods have been proposed in order to explain the working mechanism, decision behavior, and decision basis for a given trained learning model using interpretation methods or constructing interpretation models (Du et al., 2019). A local interpretation method can aid in understand the decision process and decision basis of the learning model for each particular input sample by analyzing the contribution of each dimensional feature of the input sample to the final decision outcome of the model (Guidotti et al., 2018). Classical local interpretation methods have been proposed, such as sensitivity analysis interpretation (Robnik-Šikonja & Kononenko, 2008), local approximation interpretation (Guidotti et al., 2018), gradient back-propagation interpretation (Shrikumar et al., 2016), and class activation mapping interpretation (Zhou et al., 2016).

The class activation mapping interpretation method for interpreting DCNN models is simple to implement and computationally efficient, and the interpretation results are visually appealing and easy to understand (Chen et al., 2020). It is based on the localization ability of convolutional units, which can locate the core regions in the input samples for DCNN decision making, such as decision features in classification tasks and object positions in target detection tasks (Guidotti et al., 2018). Min

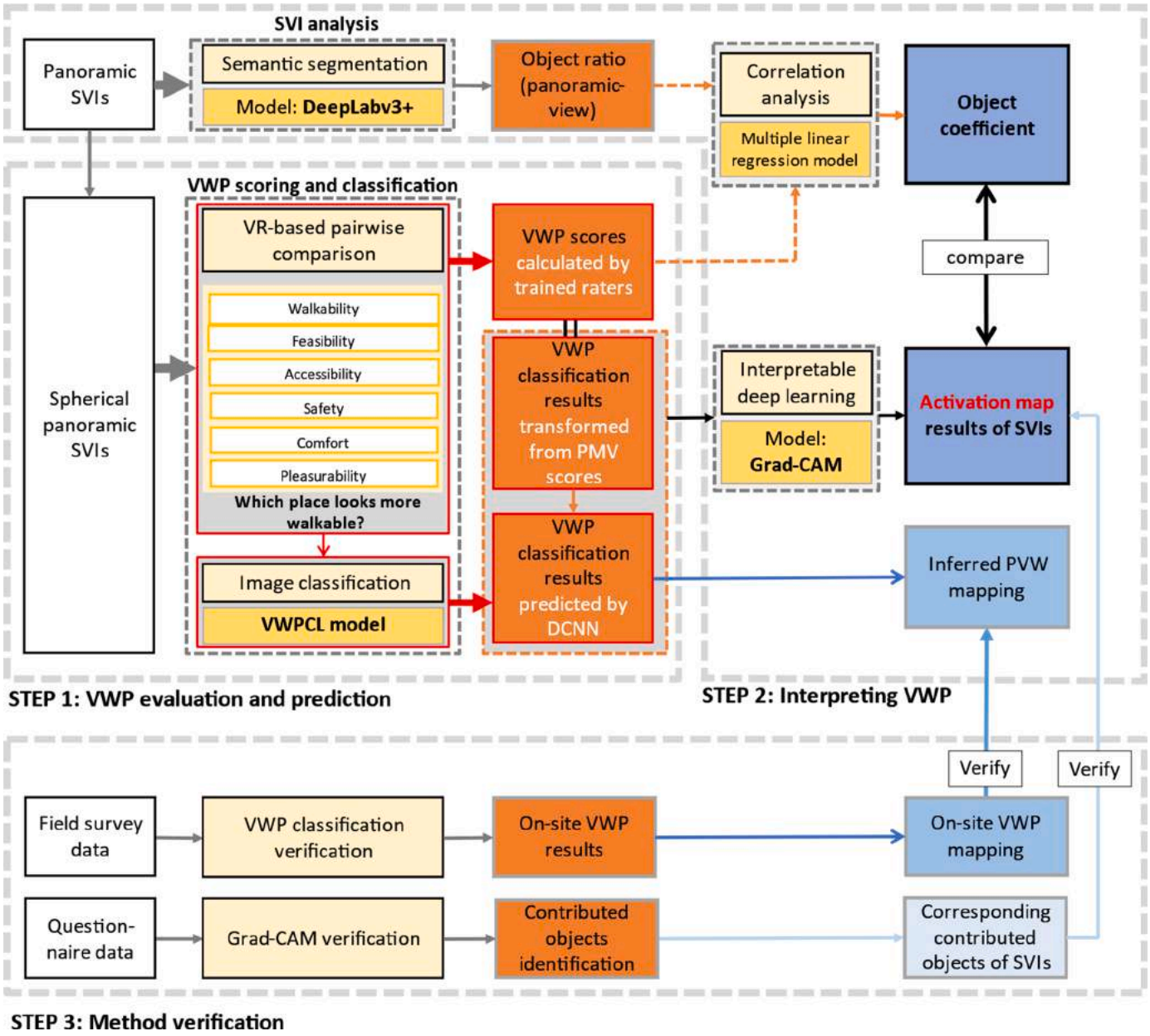


Fig. 2. Framework of the study.

et al. (2020) proposed a multi-task deep relative attribute learning network for visual urban perception attributes on the Place Pulse 2.0 dataset (Dubey et al., 2016) and adopted gradient-weighted class activation mapping (Grad-CAM) to generate visualized discriminative localization maps of six attributes. Oki & Kizawa (2021) discussed evaluation results for street attractiveness by comparing between gaze analysis calculated by eye-tracker and activation heatmaps generated by Grad-CAM from a CNN model that can estimate street-level visual impression. The above studies demonstrated that interpretable deep learning and Grad-CAM have great potential for validating models and aiding analysis in street visual perception studies, but a gap remains for VWP measurement.

3. Materials and method

3.1. Framework

In this study, a VR panoramic-based deep learning framework was developed for measuring pedestrian willingness to walk in visual perception and for quantifying and visualizing the contributed visual street-level features Fig. 2). This framework is a three-stage process: ((1) VWP evaluation and prediction, (2) interpreting VWP, and (3) method verification. This framework not only automatically evaluates VWP at large scales using easily accessible panoramic street views, but also interprets and validates the evaluation models so that stakeholders can trust model decisions and tailor the walkability detection, design, or renewal of streets based on built environment characteristics. Specifically, a VWPCL model was first developed and trained on human ratings of 360-degree panoramic SVIs in a VR system to predict street-level VWP in six categories (walkablity, feasibility, accessibility, safety, comfort, and pleasurability). This approach is able to predict these six VWP categories in a new region. Second, to investigate and interpret "what visual elements impact visual walkability perception," both statistical analysis of overall data and "visual explanation" analysis of individual data were carried out. For the former, a linear regression model was combined with segmentation of the VSIs into 19 object categories and used to determine the degree of correlation of various objects with one of the six VWP categories. For the latter, Grad-CAM, an interpretable deep learning model, was introduced to assist in identifying and visualizing elements of the street built environment that contribute to different VWP categories. Third, to verify our method, we collected onsite VWP data through a field survey on the campus of an anonymous university and questionnaire data on identification of contributing objects from the VR-based VWP evaluation. These data were used to verify the developed VWPCL model and the Grad-CAM model, respectively.

Three data sources were used in this study: (1) the VR panoramic SVIs with VWP scores were used as the VRVWPR dataset to train the VWPCL model to predict VWP scores, (2) the SVIs and field survey data of on-site VWP were used to predict and validate VWP scores based on the proposed VWPCL model, and (3) the questionnaire data of contributing objects identification from VR-based VWP evaluation were used to validate visual explanation results based on the Grad-CAM model.

3.2. VR panoramic SVI-based and VRVWPR dataset

We made VR panoramic SVIs and created a VR Visual Walkability Perceptual Rating (VRVWPR) dataset containing 2642 panoramic SVIs with VR-based human VWP ratings of urban streets. Fig. 3 shows the workflow for constructing the VRVWPR dataset. We first collected SVIs from 360-degree cameras and street view services, sphere-mapped the SVIs with a game engine, and finally collected perceptual ratings of visual walking ability through VR-based random pairwise comparisons by trained raters.

Table 1 lists the image sources in the dataset, covering cities of different regions and scales. Following Gong et al. (2018), we randomly captured 2642 panoramic SVIs from street view services. The image size for both acquisition methods is 5376×2688 pixels. These blended images indicate the diversity and heterogeneity of the dataset.

After sphere mapping in the game engine, the SVI data were stored in a database and connected to an HMD device in preparation for the perceptual scoring of visual walking ability. In HMD devices, the raters are presented with two panoramic SVIs randomly selected from the entire database, and they could use VR controllers to roam through 360-degree panoramic SVIs. The raters were then asked to choose one of the SVIs or identify them as "equal" in response to one of the six questions in Fig. 4 without limitation of decision time and image-switch chances: "which place looks more walkable?" and "which place looks more X for

Table 1
Statistics of image data in the VRVWPR dataset.

Country	City	No. of Images
United States	New York	365
United States	Los Angeles	318
United Kingdom	London	302
France	Paris	405
Germany	Berlin	335
Japan	Tokyo	475
Japan	Osaka	442

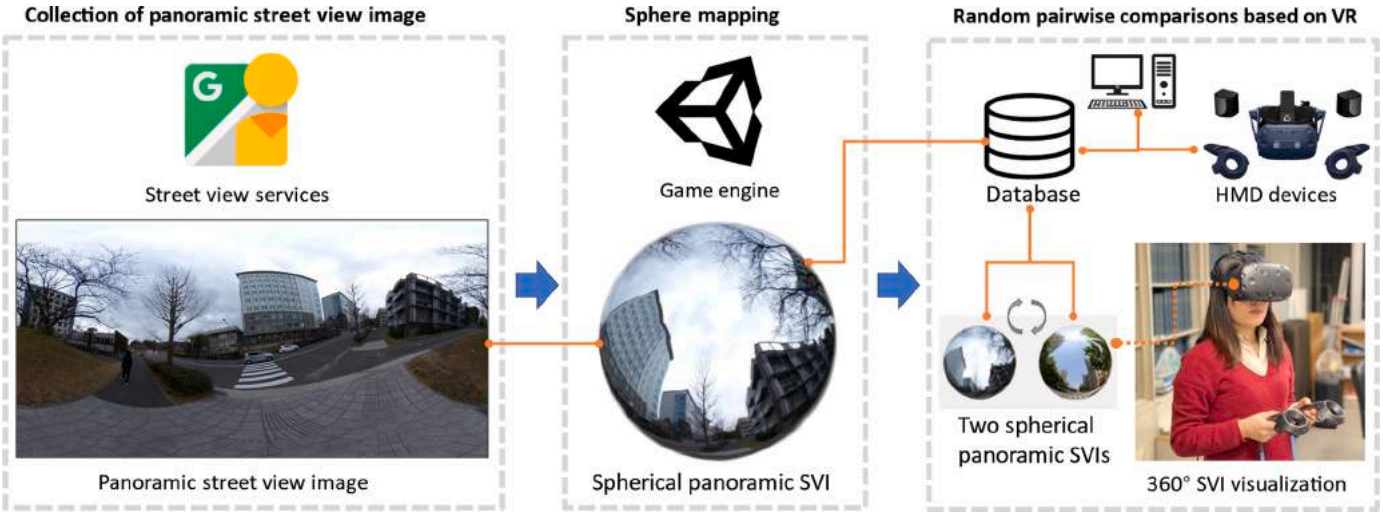


Fig. 3. Workflow of the VRVWPR dataset construction.

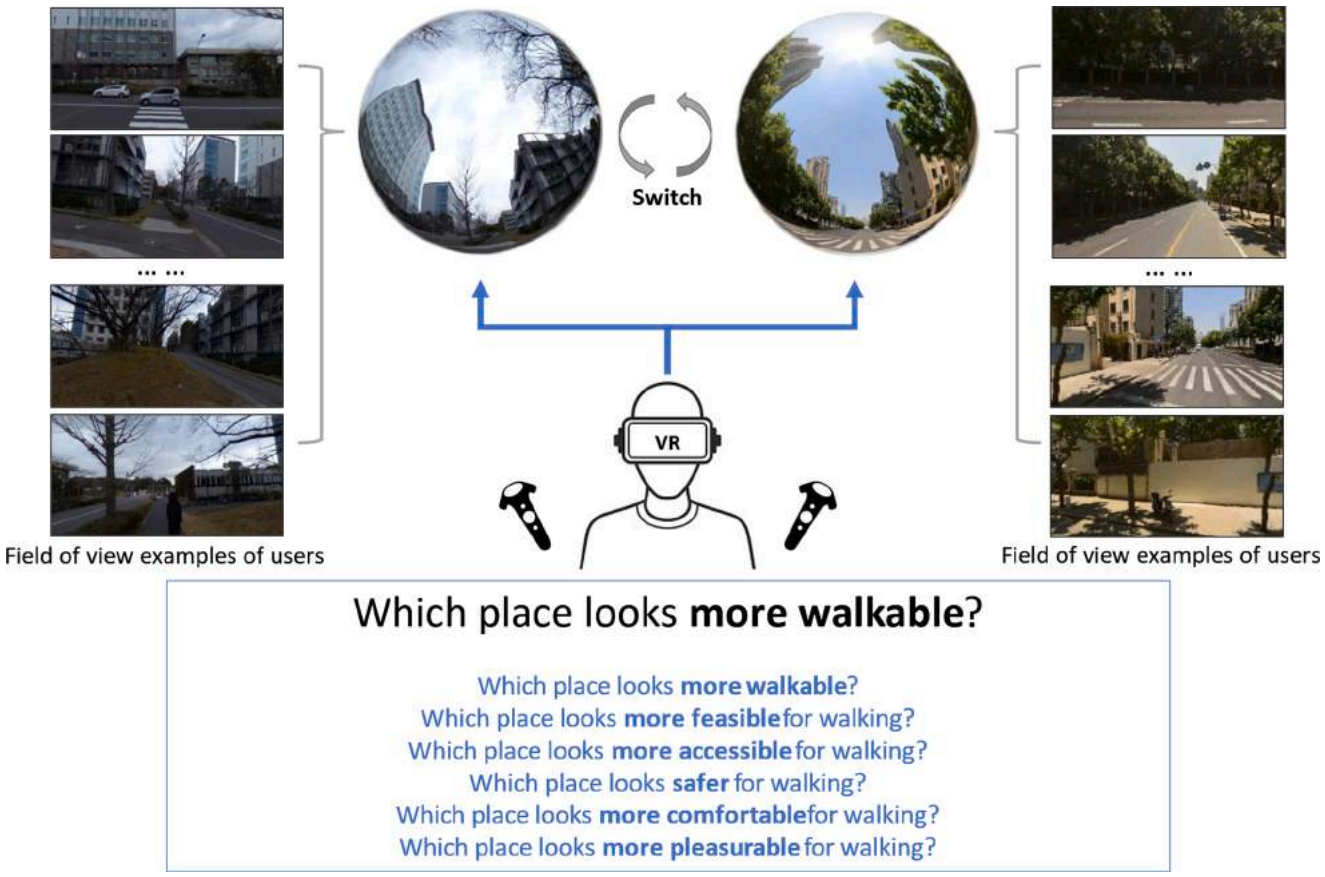


Fig. 4. Trained raters with HMD devices were asked to choose one of two images in response to one of six questions for the VWP categories.

walking?" where the X can mean "feasible," "accessible," "safe," "comfortable," or "pleasurable." These six questions correspond to the six VWP categories. The ratings were based on direct perception and no difference threshold was applied to decide on the equal rating value.

Thirty raters (15 female and 15 male), all designers with a background related to urban design education, were invited to participate in the perceptual scoring of visual walkability. They made a total of 20,549 pairwise comparisons and each participant rated nearly 685 pairs in the process. The ratio of equally rated pairs in the overall decisions is 10.1%. None of the raters reported simulator sickness. To prevent the results from being influenced by subjective bias, the raters were first trained on the main visual factors in the hierarchy of walking needs (Fig. 1) to establish a consensus on the criteria, and then they performed scoring separately.

Following Salesses et al. (2013) and Zhang et al. (2018), to convert the results of pairwise comparison of images into VWP scores of a single image, VWP scores with respect to a specific question, S_i, can be calculated by using Eqs. (1), (2), and (3), and then scaled to a range of 0 to 10. Fig. 5 depicts four image examples with their scores for each of the six VWP categories.

$$B_i = b_i/(b_i + e_i + w_i), \tag{1}$$

$$W_i = w_i/(b_i + e_i + w_i), \tag{2}$$

$$S_i = \frac{10}{3}\cdot\left(B_i + \frac{1}{b_i}\sum_{n=1}^{b_i} B_n - \frac{1}{w_i}\sum_{m=1}^{w_i} W_m + 1\right), \tag{3}$$

where B_i and W_i are the better rate and worse rate of image i along a

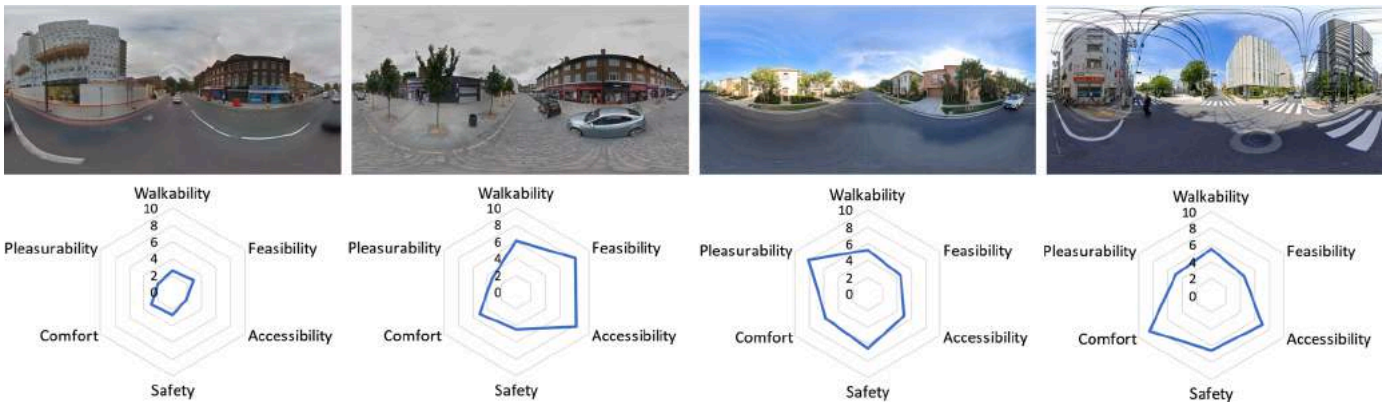


Fig. 5. Image samples from the VRVWPR dataset with their perceptual score of the six categories.

specific VWP categories, respectively; S_i is the corrected score for image i in a specific VWP category; b_i, w_i, and e_i are the number of times that image i was selected, not selected, or judged to be equal to another image in the comparisons for a specific question, respectively; and the first and second sums extend over n and m, respectively.

3.3. VWP evaluation and prediction

Inspired by image classification technology based on DCNNs, combined with the VRVWPR dataset, we designed a VWPCL model for VWP classification to predict pedestrians' VWP for an SVI in six categories. Fig. 6 illustrates the workflow of the proposed VWPCL model.

In the VRVWPR dataset, each SVI has a corresponding VWP score in six categories. To accomplish the image classification task, we first transform the VWP scores of each category in the VRVWPR dataset according to the score bands, labeled as low (<3), medium (3–7), and high (>7). Table 2 shows the image classification statistics of the VRVWPR dataset in six categories. Then, we used 80% (2113 images) and 20% (529 images) of these data as training and testing subsets of the VWPCL model, respectively.

The VWPCL model is inspired by the DenseNet architecture and multitask learning. DenseNet, a DCNN, slows gradient disappearance, enhances feature transfer, and reduces the number of parameters, achieving excellent performance in image classification tasks. Multitask learning, a type of migration learning, can obtain additional useful information by mining the relationship between multiple tasks and has better model generalization ability. The proposed DCNN and multitask learning network of the VWPCL model automatically learns the shared deep feature representations among the six VWP classification tasks and can support classification tasks with small task samples.

3.4. Interpreting VWP

To interpret and determine which visual elements may be associated with a place's VWP, a correlation analysis based on a multiple linear regression model was first used to determine, in general, which visual elements of the street built environment are favored for a place to be perceived as, for example, walkable and safe. Then, the Grad-CAM technique based on interpretable deep learning was used to highlight and visualize the salience of the features for each VWP category.

3.5. Correlation analysis between the object ratio of SVIs and VWP scores

Fig. 7 shows the workflow of the correlation analysis between the object ratio of SVIs and VWP scores. The correlation analyses were conducted separately for each of the six VWP categories. First, VWP scores were obtained from the VRVWPR dataset. Meanwhile, to calculate the percentage of elements in the built environment of the street, the DeepLabv3+ model based on semantic segmentation with the Cityscapes dataset was used. It can extract pixel-level semantic information of 19 physical components, namely, vegetation, roads, sidewalks, terrain, building, wall, fences, sky, cars, trucks, buses, trains, motorcycles, bicycles, poles, traffic lights, traffic signs, people, and riders. The area ratio of the 19 physical component segments was obtained and interpreted as color groups on each panoramic SVI with 81.3% of mean intersection-over-union. Finally, stepwise multiple linear regression models (Eq. (4)) were employed to investigate the relationship between the 6-category VWP scores and 19 physical components. Because stepwise regression allows regression models to be constructed from a set of candidate variables, the system automatically identifies the influential variables. This allows for the elimination of independent variables lacking significance and the creation of an "optimal" regression equation for data with many variables that may not be completely independent of each other.

$$y_i = \beta_0 + \beta_1 + x_i + \ldots + \varepsilon_i,\ i = 1, \ldots,\ n \tag{4}$$

where y_i is the response variable; x_i represents the regression variables; $\beta_1, \beta_2, \ldots, \beta_n$ are partial regression coefficients; ε_i is an error term; and the subscript i indexes a particular observation.

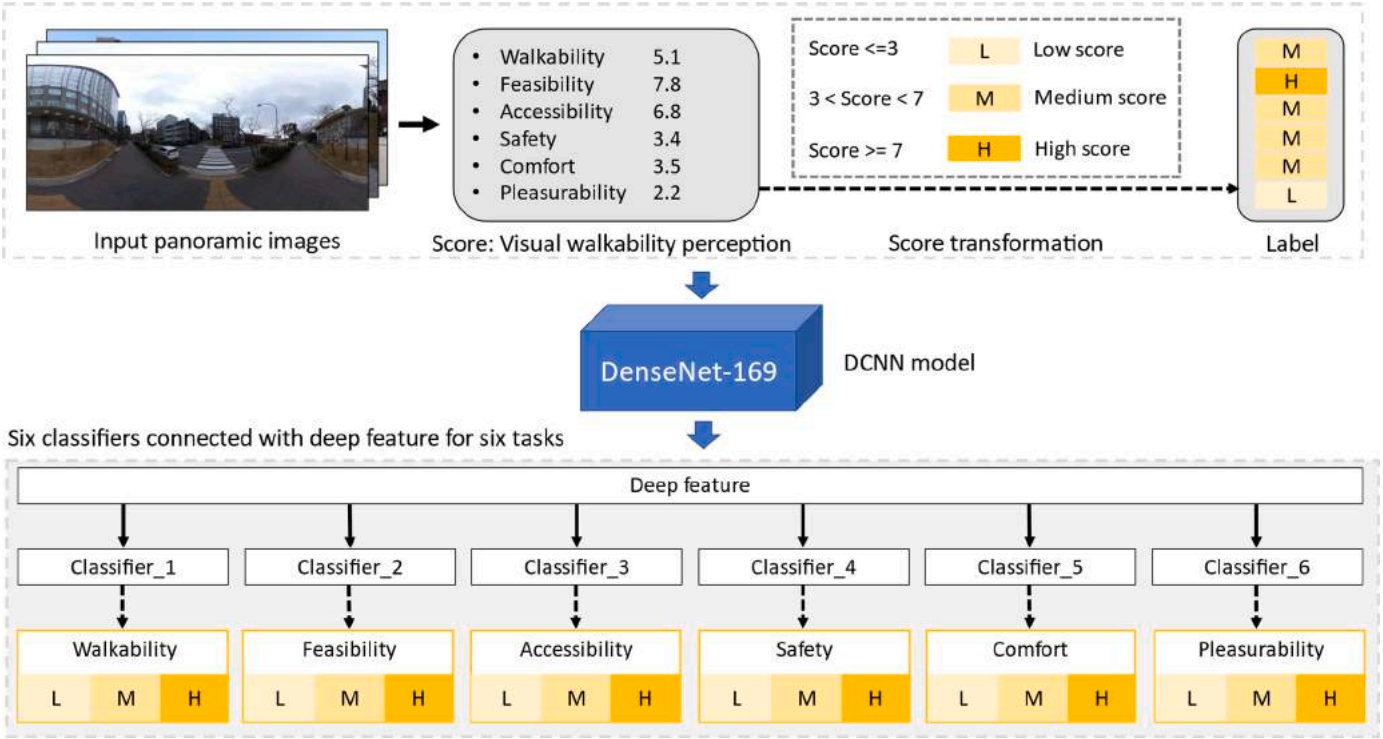


Fig. 6. Workflow of the developed VWPCL model.

Table 2
Image classification statistics of the VRVWPR dataset in six categories (H: high score; M: medium score; L: low score).

	#Walkable	#Feasibility	#Accessibility	#Safety	#Comfort	#Pleasurability
H	702	660	742	408	542	672
M	914	664	1054	980	1186	1072
L	1026	1318	846	1254	914	898

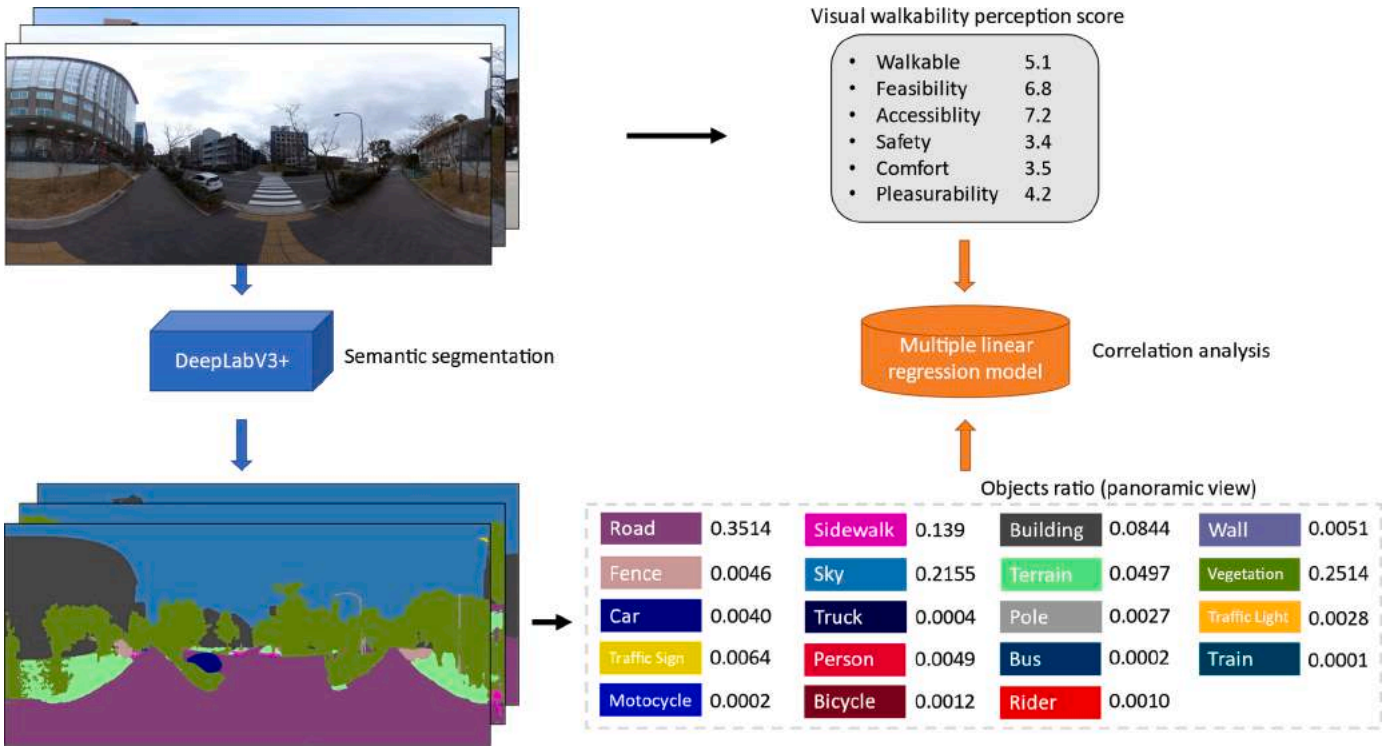


Fig. 7. Workflow of the correlation analysis between the object ratio of SVIs and VWP scores.

3.6. *Interpretable deep learning for VWP results*

An interpretable deep learning technique, Grad-CAM, was employed to determine the VWPCL model transparency by using the layers and extracted features of the trained model, as a visual explanation method for VWP results. Grad-CAM is a classical class activation mapping method that combines gradient information with feature mapping for gradient weighting (Fu et al., 2020). Given an input panoramic image sample, on the basis of the pre-trained VWPCL model, Grad-CAM first calculates the gradient of the target class with respect to each feature map in the last convolutional layer and globally averages the gradient pooling to obtain the importance weight of each feature map; then, the weighted activation of the feature map is calculated based on the importance weight to obtain a gradient-weighted class activation map for locating the important class-discriminative regions in the input samples with class-discriminative properties (Selvaraju et al., 2017). The highlighted regions of the input image are generated by the feature maps in the final convolutional layer that hold the spatial information from captured visual patterns, implying that these locations receive the most attention from the model during the classification process. Fig. 8 is the workflow of interpretable deep learning for VWP results using Grad-CAM. In the output activation heat map, warmer color of the

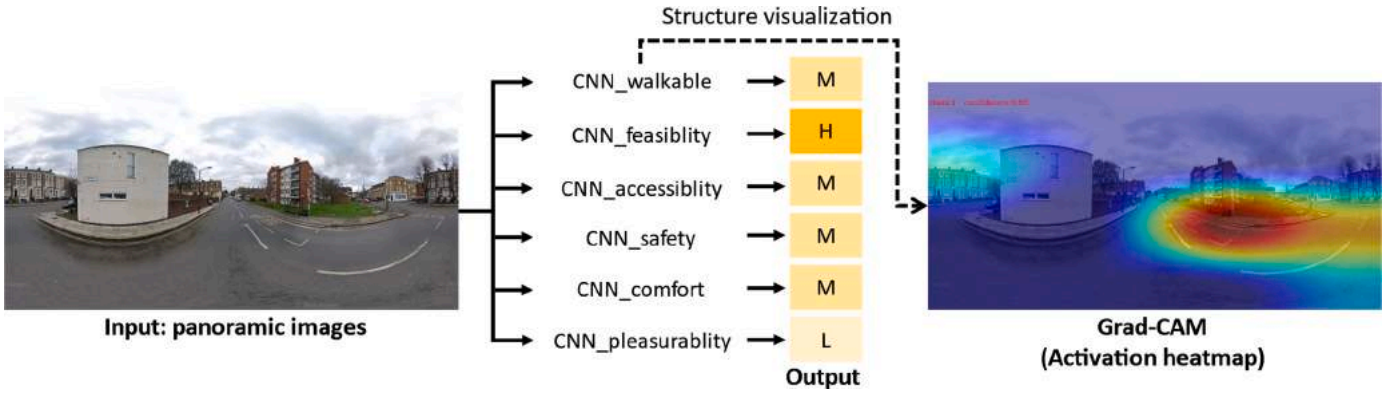


Fig. 8. Workflow of interpretable deep learning for VWP results.

overlay image indicates that a pixel is more discriminative.

3.7. Method verification

The deep learning models needing verification in our method are the proposed VWPCL model and the Grad-CAM model. Twenty volunteers were recruited to participate in the verification, and the characteristics of these volunteers are shown in Table 3.

For the VWPCL model verification, to predict and verify the VWP scores of new urban areas, we used the campus of an anonymous university as an example and collected SVIs and field survey data within this area. When acquiring panoramic SVIs with the same image size in the VRVWPR dataset from street view services, we set sampling points at 50-m intervals on each street and collected a total of 906 panoramic images as on-site validation dataset. However, it is very hard to apply pairwise comparison in on-site auditing with 906 sampling points for the field survey. Volunteer direct-scoring based on field survey is often used for perception auditing (Li et al., 2020; Li et al., 2022; Tang & Long, 2019) is a cost-effective and efficient way to verify the auditing discrepancy in VR and real-world (Kim & Lee, 2022). It has advantages in freely changing the on-site validation dataset size while controlling the validation time and ensuring a certain level of accuracy. Therefore, the 20 volunteers were first trained on the same evaluation criteria (Fig. 1) with the previous pairwise comparison of VRVWPR dataset making, and then were asked to perform an on-site assessment of the six VWP categories through a field survey at each sampling site within the campus area based on their visual perception of walking, with each item scored on a scale from 0 to 10. The on-site VWP scores were averaged across participants.

To validate the visual explanation results from the Grad-CAM model, 50 images from the VRVWPR dataset were randomly selected and used as the data source for the questionnaire. The 20 volunteers were asked to evaluate the VWP based on the VR panoramic SVIs and to identify in the questionnaire the elements of the street built environment that visually serve to increase their willingness to walk there.

4. Experiment and results

4.1. VWP evaluation and prediction results

The classification results of the VWPCL model based on the VRVWPR dataset are shown in Table 4, and Fig. 9 presents the normalized confusion matrices for the six VWP classification tasks. Fig. 10 shows image samples from Osaka that were predicted to have high scores, medium scores, and low scores for VWP. The model achieved an overall accuracy of 85.4% for the VWP classification in the six categories. However, some subcategories had a lower accuracy of around 70%, such as the low subcategory under accessibility. Furthermore, the overall lowest accuracy in the classification tasks was 76.8% in the accessibility category, but this is still acceptable considering the dataset's diversity and complexity. In addition, we noted that some subcategories were misclassified as each other more easily, such as the medium and low score subcategories in "safety." This may be due to the presence of some similar streetscape elements such as low greenery and high traffic flow in the low and medium score results of "safety"; however, some visual elements, such as graffiti and garbage, that distinguish low and medium score results usually have a small pixel share in the SVIs.

Table 3
Characterization of individuals in the questionnaire survey.

Demographics	Group	Frequency	Percent
Gender	Male	10	50%
	Female	10	50%
Education	Bachelor	4	20%
	Master	8	40%
	Doctor	8	40%
Major	Architecture	6	30%
	Urban planning	6	30%
	Landscape planning	4	20%
	Environment engineering	4	20%

4.2. Results of VWP interpretation

4.2.1. Factor identification results of correlation analysis

Fig. 11 shows the results of the stepwise multiple regression analysis between the physical components in segmented panoramic SVIs and the perception scores of the VRVWPR dataset, where independent variables with significant positive (blue bars) or negative (orange bars) effects on each VWP category are ranked and listed. A total of six stepwise regression models were constructed, corresponding to the six VWP categories (walkability, feasibility, accessibility, safety, comfort, and pleasurability) with physical components. Stepwise regression adds or excludes all the regressors to propose only significant models. In this study, each VWP category had several possible regression models, but only the significant models with the maximum number of explanatory variables are discussed. The degree of fitting (R^2 values) for each of the six models was 0.594, 0.457, 0.408, 0.536, 0.467, and 0.593, respectively, which means that the listed physical components in Fig. 11 can explain about half of the variation in each VWP category. All the variance inflation factor values in these models are less than 5, and the Durban–Watson values are all around 2, which means that cointegration is not a problem.

As shown in Fig. 11, "vegetation," "sidewalk," "terrain," and "road" were the top four explanatory variables in the other five VWP categories except for "feasibility." Among them, "vegetation" is more sensitive to "safety" and "pleasurability," while "sidewalk" is more sensitive to "accessibility." These results are consistent with the previous proposition that urban greenery brings a sense of safety, affects pedestrians thermal comfort, and improves street attractiveness (Alfonzo, 2005; Ashihara, 1983; Bosselmann et al., 1999; Jiang et al., 2014; Quercia et al., 2014) and a wider pedestrian space increases street walking capacity and makes it more comfortable for pedestrians to pass through (Frackelton et al., 2013).

The objects that slightly contribute to the other five VWP categories varied. For instance, the area ratio of "person" and "bicycle," which helps to make a street lively (Zhang et al., 2018) was slightly positively correlated with the "walkability" and "pleasurability" scores. "Truck" and "motorcycle," which bring traffic noise and pose potential traffic safety hazards, have a negative effect on "walkability," "feasibility," "accessibility," "safety," and "comfort." "Traffic lights" and "traffic signs" play a positive role in "feasibility," "accessibility," "safety," and "comfort." These two elements are important components of the street infrastructure for pedestrians, influencing pedestrian willingness to walk and enhancing traffic safety and the perception of the neighborhood environment (Li et al., 2020). In addition, it is noteworthy that "fence" plays a key role in countering "feasibility" among the most important nine related elements, which aligns with the fact that a number of actual or perceived barriers are particularly important in impeding the initiation of walking at the street level (Alfonzo, 2005).

4.2.2. Interpretable results for VWP using Grad-CAM

Fig. 12 shows Grad-CAM result samples with high, medium, and low scores in the six VWP categories, where the contributing visual regions are highlighted in the image. Similar to the results of the stepwise regression described in the previous section, greenery, such as street trees and lawns, was activated in the samples with high score results for walkability and high score results for pleasurability, indicating the stronger positive influence of "vegetation" and "terrain" on "walkability" and "pleasurability." Streets without sidewalks were activated in low score results for accessibility, and barricades were activated in medium score results for accessibility, indicating the important influence of "sidewalk" and "fence" on "accessibility." Crosswalks and some

Table 4
Classification accuracy, precision, recall, and F1 score of the VWPCL model (H: high score; M: medium score; L: low score).

Overall accuracy	VWP category	Type	No. of Test samples	Precision	Recall	F1 score	Accuracy
85.4%	Walkability	H	140	85.1%	85.7%	0.85	87.1%
		M	183	90.0%	86.5%	0.88	
		L	205	86.2%	88.6%	0.87	
	Feasibility	H	132	90.4%	92.8%	0.91	91.5%
		M	132	89.3%	87.3%	0.88	
		L	264	94.7%	94.5%	0.95	
	Accessibility	H	148	85.0%	67.9%	0.76	76.8%
		M	211	75.3%	79.0%	0.77	
		L	169	70.1%	83.0%	0.76	
	Safety	H	81	86.0%	62.3%	0.72	80.6%
		M	196	81.4%	72.6%	0.77	
		L	251	74.4%	94.2%	0.83	
	Comfort	H	108	85.7%	91.7%	0.89	90.4%
		M	237	91.6%	93.2%	0.93	
		L	183	93.8%	89.2%	0.92	
	Pleasurability	H	134	88.1%	85.5%	0.87	86.2%
		M	214	85.2%	84.6%	0.85	
		L	180	85.4%	87.4%	0.86	

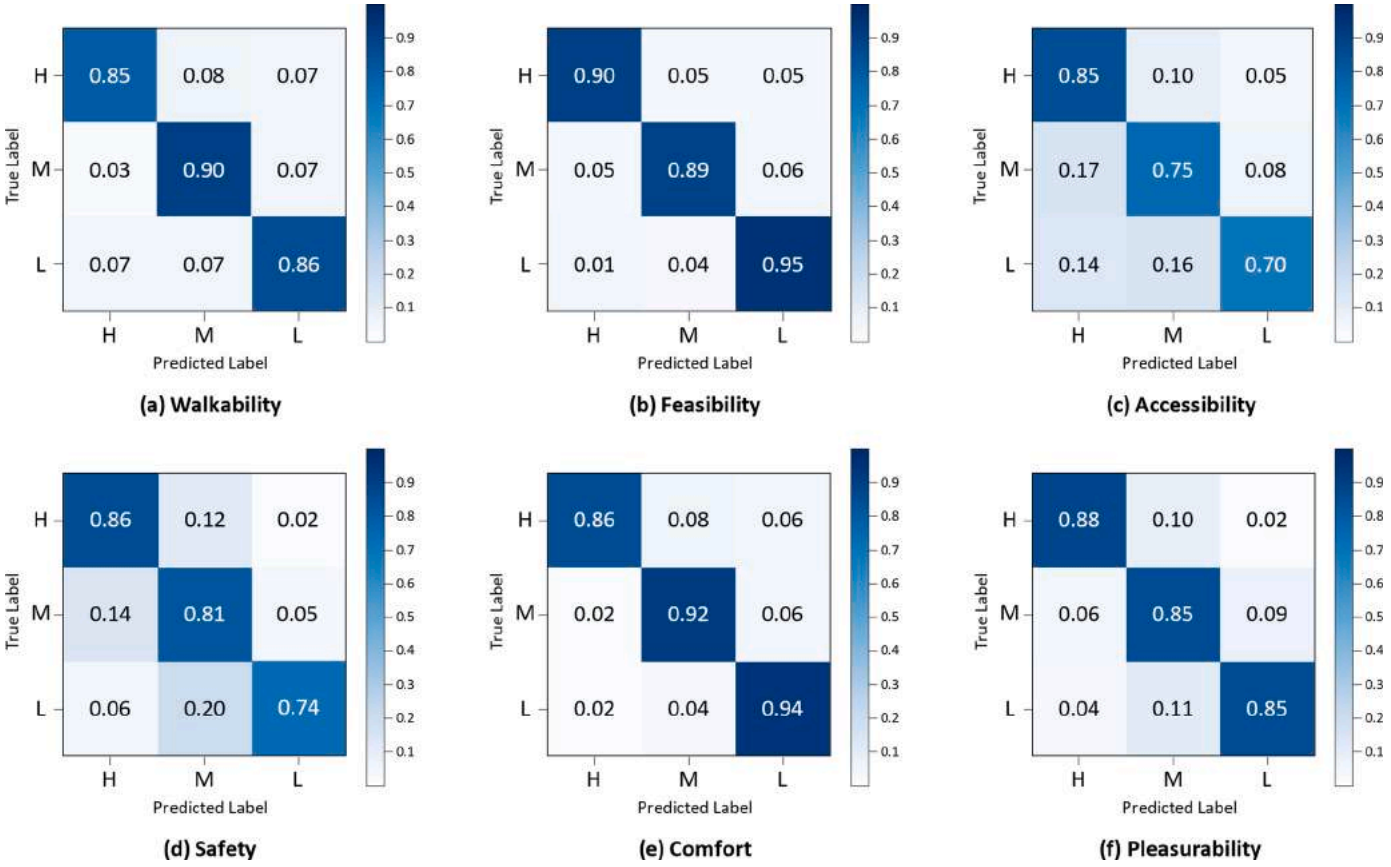


Fig. 9. Normalized confusion matrices for six classification tasks: (a) walkability, (b) feasibility, (c) accessibility, (d) safety, (e) comfort, and (f) pleasurability.

traffic signs were positive visual elements corresponding to safety attributes.

In particular, Grad-CAM activated some reasonable streetscape elements that were not categorized as one of the 19 physical components in semantic segmentation. For example, a medium score result for feasibility activates store signs, a medium-score result for comfort activates blind corridors, a high score result for comfort activates bicycle parking facilities, a low score result for safety activates buildings under construction, and a high score result for pleasure activates street furniture. These activated elements are not segmented in the semantic segmentation model, but they have an important role for different VWP categories. For example, store signs suggest street vibrancy and rich land use types that help promote walking behavior (Li et al., 2022). Blind corridors, bicycle parking facilities, and street furniture are important elements of street infrastructure to enhance pedestrian-related street quality, which are closely related to comfort and enjoyment (Alfonzo, 2005).

In addition, although the stepwise regression results show that roads have a positive effect on "feasibility," large bare roads were activated in the low score result for feasibility, indicating that appropriate street width and walking-friendly street scale are very relevant to "feasibility." This also indicates that the results of the stepwise regression cannot

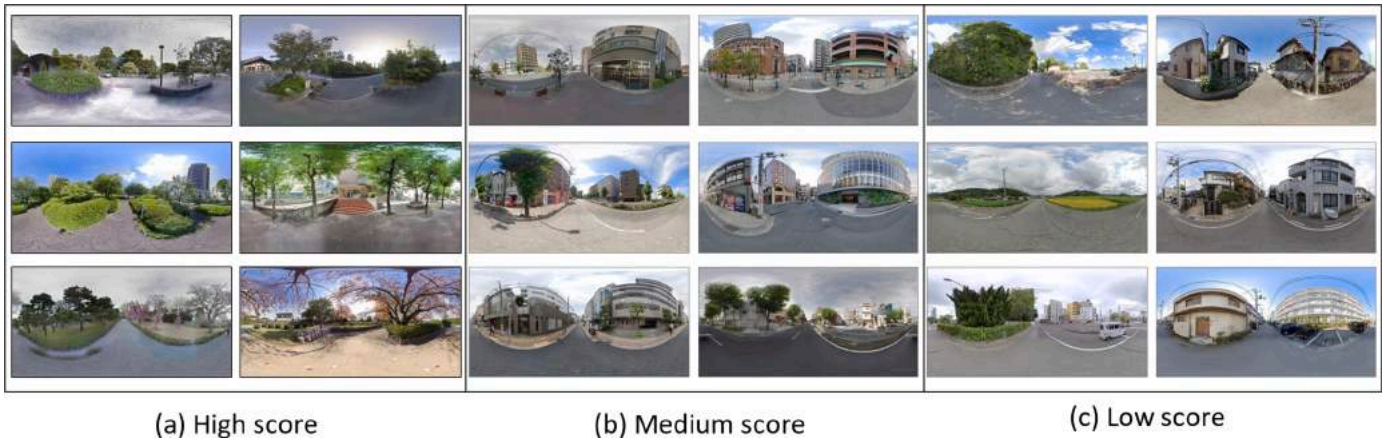


Fig. 10. Image samples from Osaka that were predicted to have (a) high scores, (b) medium scores, and (c) low scores for VWP.

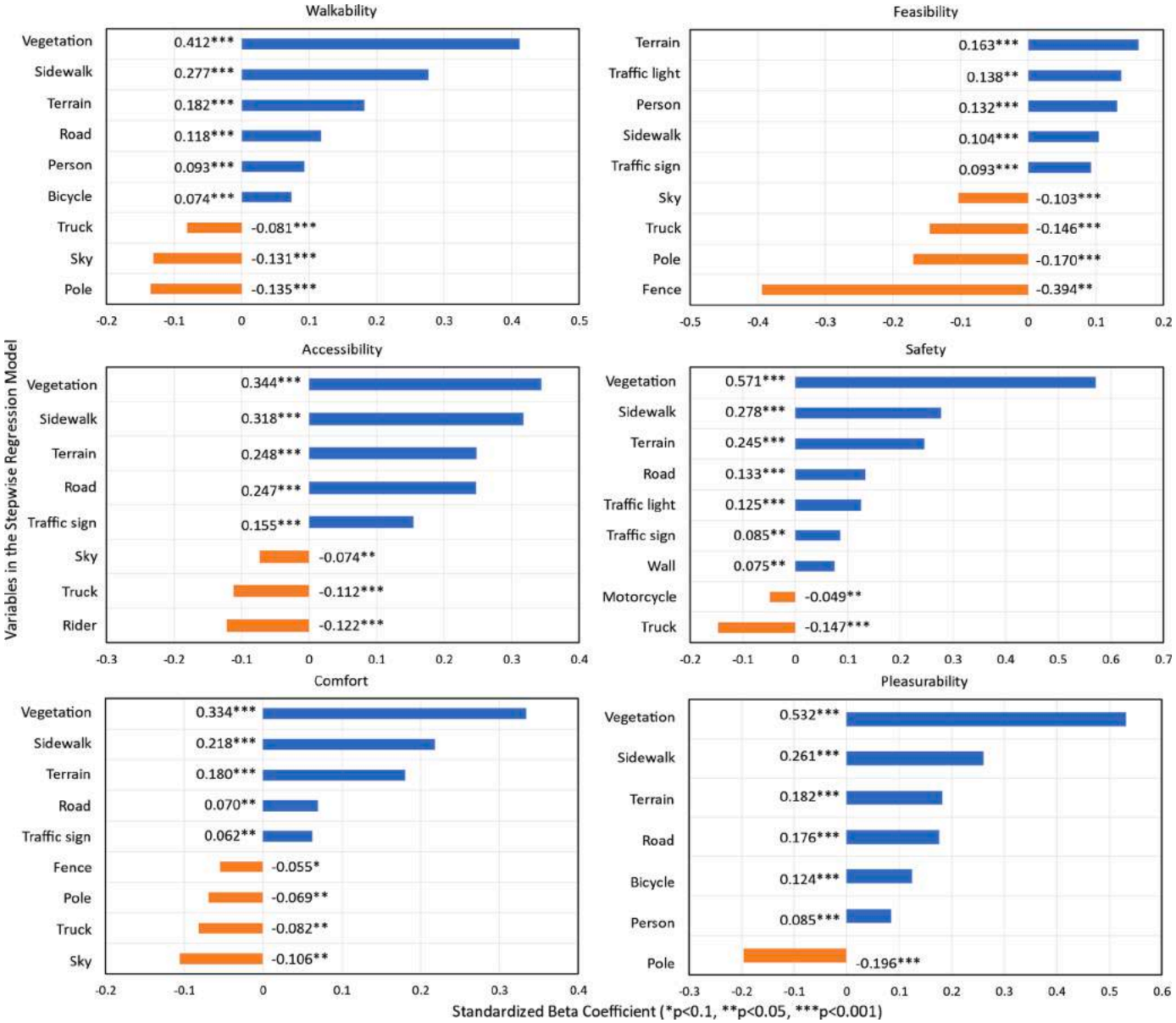


Fig. 11. Results of the stepwise multiple regression analysis between the physical components and perception scores. For each pair, independent variables with significant effects not excluded from the stepwise regression models (out of 19 physical components) are shown.

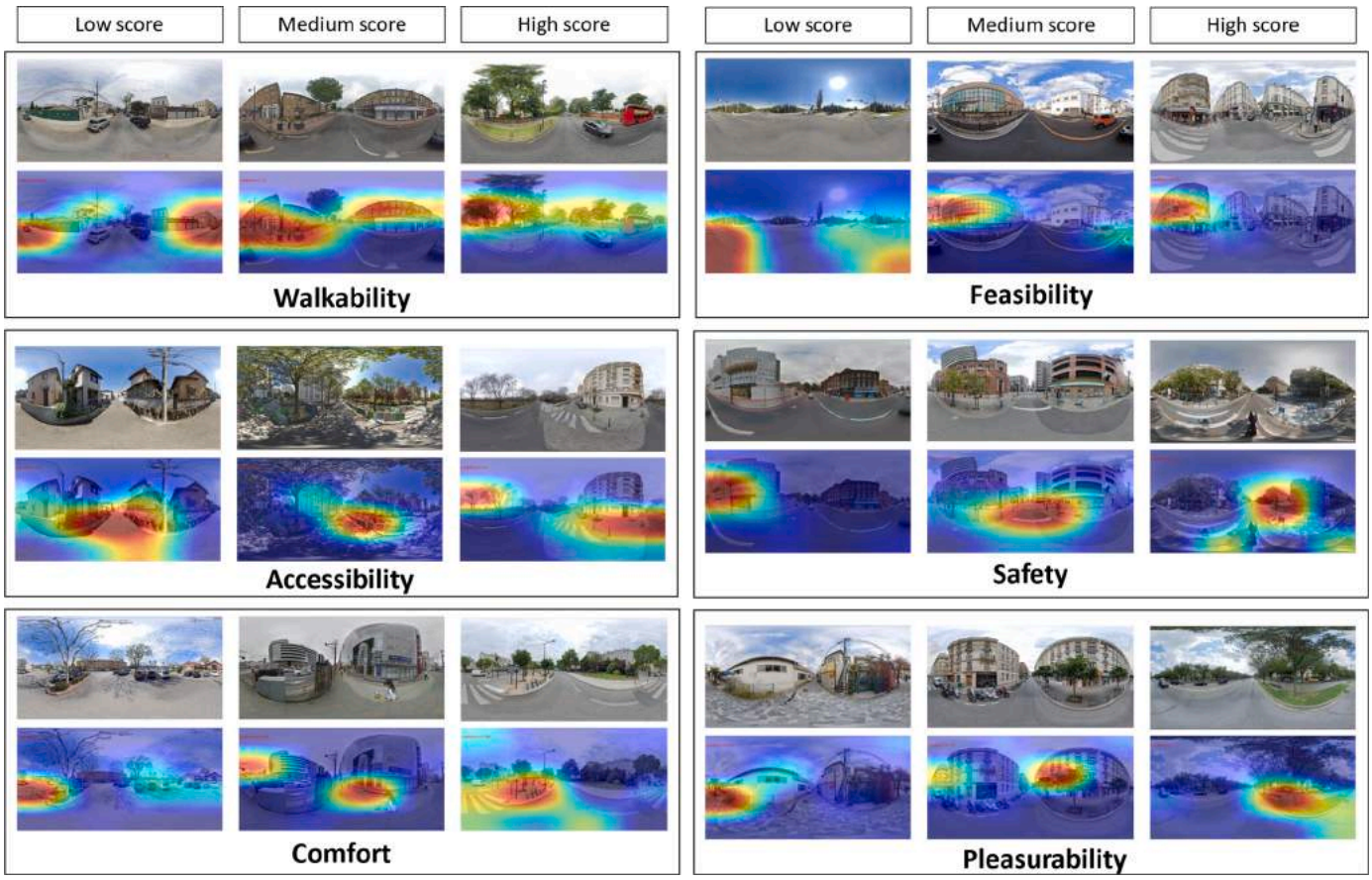


Fig. 12. Examples of Grad-CAM results for six-category VWP classification.

explain some streetscape elements related to street scale and spatial sense when only the percentage of street elements is used as the explanatory variable.

4.3. Method verification results

4.3.1. VWP classification verification based on-site auditing

To apply the proposed classification method, the campus of an anonymous university was used for an on-site verification experiment, and all the street VWP results in six categories were mapped based on the pre-trained VWPCL model. Fig. 13 presents the region maps of the on-site verification experimental area, where 906 sampling points with panoramic SVIs and on-site rating results of recruited volunteers were obtained. The pre-trained VWPCL model predicted VWP classification results in the six categories, and the results are shown in Fig. 14. Green points, yellow points, and red points represent the areas with high, medium and low scores, respectively, corresponding to the VWP categories.

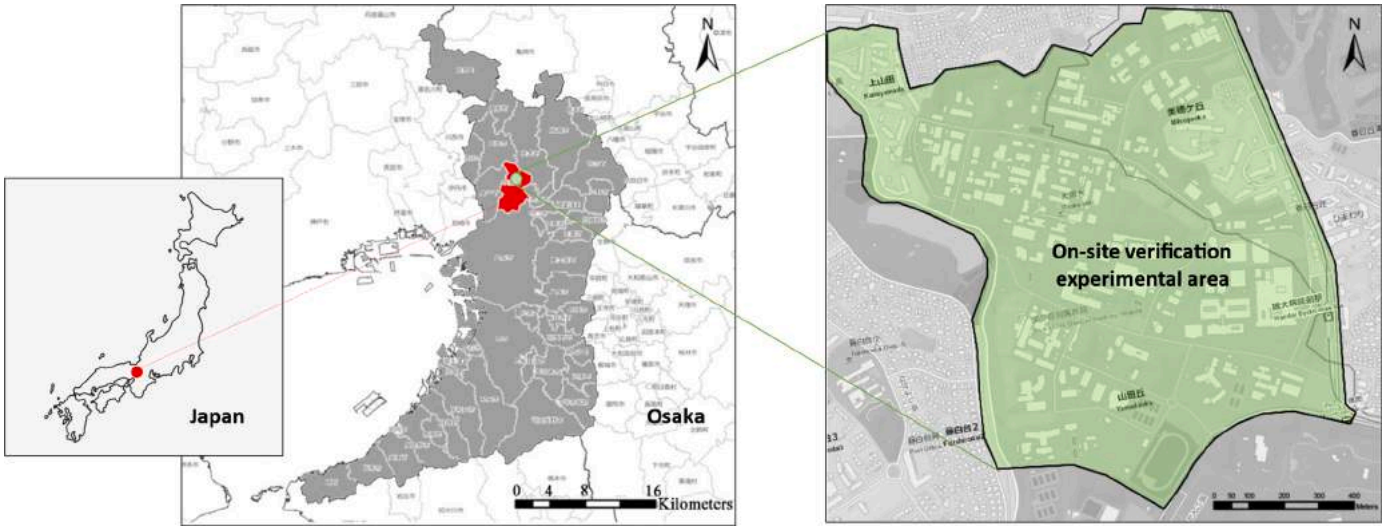


Fig. 13. Region maps of the on-site verification experimental area.

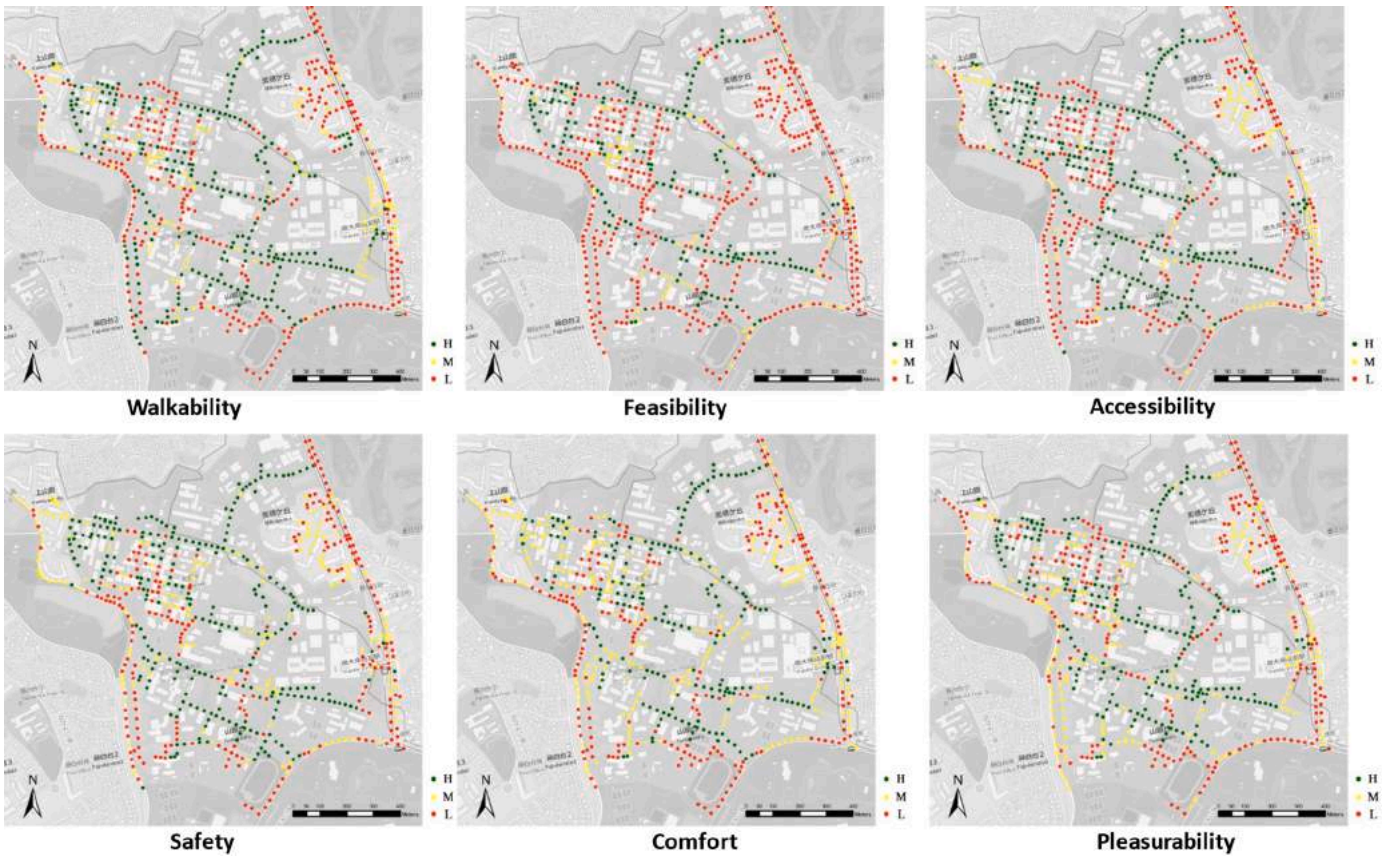


Fig. 14. Mapping the predicted results of on-site verification area for VWP using VWPCL model in six categories.

As shown in Fig. 14, the streets with high visual walkability are mainly located on internal campus roads, especially on major traffic arteries. The overall feasibility of the streets in the verification area is low, and the high-feasibility streets are sporadically distributed within the campus. Streets with high accessibility are concentrated on major traffic arteries and in dense road network areas in the northwest of the campus. The overall safety of the streets within the campus is high, while the streets with low safety are concentrated on the municipal roads on the east and south sides of the campus, which have a tram overpass and high traffic flow. The distribution of street comfort is more balanced in the verification area, but the high-scoring areas are still concentrated in the interior of the campus. In addition, the overall street pleasurability is relatively higher in neighborhoods with low building density.

Using the on-site auditing results by the volunteers as a benchmark, the proposed VWPCL model can achieve moderate accuracy (81.9% overall, and 88.5%, 82.6%, 71.6%, 77.1%, 82.8%, and 88.5% for walkability, feasibility, accessibility, safety, comfort, and pleasurability, respectively). The classification results of the VWPCL model in the on-site verification experimental area are shown in Table 5, and Fig. 15 presents the corresponding normalized confusion matrices for the six VWP classifications. The overall classification results are similar to those in Section 4.1, with a slight decrease in accuracy. This indicates a small

Table 5
Classification accuracy, precision, recall, and F1 score of activity-based model in VWPCL model for image samples from the campus of an anonymous university (H: high score; M: medium score; L: low score).

Overall accuracy	VWP category	Type	No. of test samples	Precision	Recall	F1 score	Accuracy
81.9%	Walkability	H	241	83.1%	81.5%	0.82	88.5%
		M	313	90.8%	84.4%	0.88	
		L	352	91.6%	95.6%	0.94	
	Feasibility	H	226	82.2%	80.3%	0.81	82.6%
		M	228	79.7%	70.3%	0.75	
		L	452	86.1%	93.5%	0.90	
	Accessibility	H	254	78.1%	71.4%	0.75	71.6%
		M	361	70.0%	69.2%	0.70	
		L	290	66.8%	74.1%	0.70	
	Safety	H	400	83.1%	91.7%	0.87	77.1%
		M	336	81.9%	71.8%	0.77	
		L	160	66.4%	70.1%	0.68	
	Comfort	H	286	79.2%	83.0%	0.81	82.8%
		M	307	83.1%	78.8%	0.81	
		L	313	86.3%	86.8%	0.86	
	Pleasurability	H	230	88.0%	88.2%	0.88	88.5%
		M	368	89.8%	86.2%	0.88	
		L	308	87.7%	91.7%	0.89	

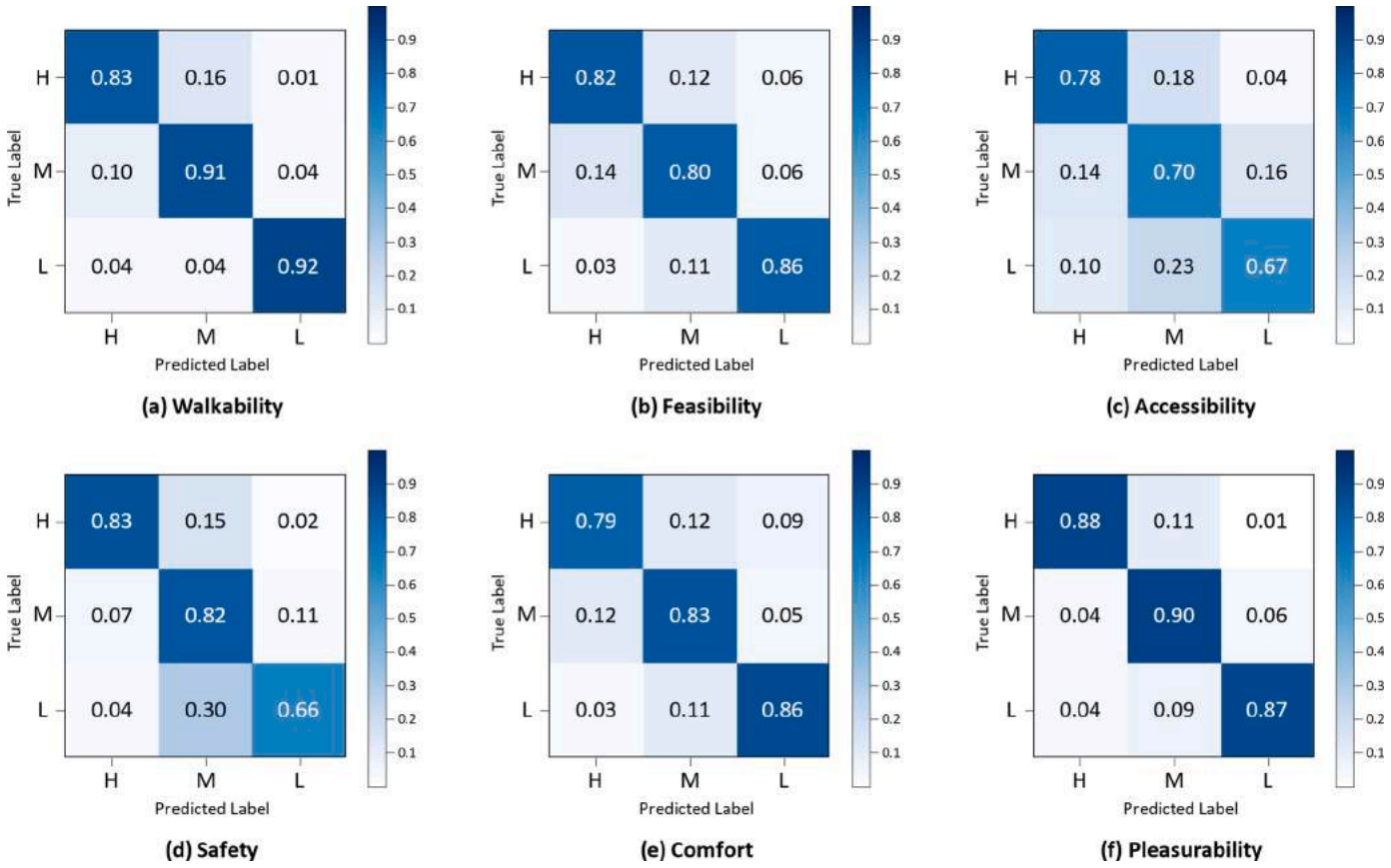


Fig. 15. Normalized confusion matrix of sampling points for the campus of an anonymous university in six-category VWP classification.

visual bias between the VR-based panoramic street map audit and the field research review, but within an acceptable range.

4.4. Grad-CAM verification: questionnaires for identifying contributing objects

Fig. 16 shows samples of Grad-CAM results for high, medium, and low scores in six-category VWP classification in the experimental area for on-site verification. Similar to Fig. 12, greenery is a frequently activated element in the medium and high score results. In addition, some streetscape elements, such as viaducts and parking signs, which are related to traffic flow and noise but outside of the 19 physical components, are reasonably activated in the low walkability score result and the low accessibility score result.

To verify whether the visual elements identified by Grad-CAM with large contributions of perceptual attributes are correct, we collected and counted questionnaire results of the 20 volunteers for 50 sampled SVIs as a benchmark for the Grad-CAM activation map results. For a VWP category in an SVI, the Grad-CAM validation results of this SVI were respectively recorded as fully consistent, partially consistent, and fully inconsistent if greater than 80%, between 20% and 80%, and less than 20% of volunteers perceived the contributing streetscape elements to be consistent with the activation map. Table 6 presents the statistical results for 50 images in Grad-CAM verification. The contributing objects in the questionnaire and activated areas in the activation heat map that were totally inconsistent for the six VWP categories were 14%, 12%, 26%, 22%, 16%, and 10%, respectively. Pleasurability had the highest percentage that was totally consistent (76%), while feasibility had the lowest percentage (48%). The average totally consistent, partially consistent, and totally inconsistent rates of the six-category VWP were 60.7%, 22.6% and 16.7%, respectively. This indicates that the activation heat map of Grad-CAM has a moderately high ability to interpret the classification prediction results of the VWPCL model, and can correctly explain most of the contributing streetscape elements.

5. Discussion

Street walkability has been a widely discussed topics in recent years, yet few studies have discussed the measurement of VWP based on SVI auditing for perception ratings and the streetscape elements that may contribute to VWP. A few studies have applied VR to panoramic SVIs to achieve immersive ratings and used Grad-CAM as an aid to validate deep learning classification models and regression models.

The advantages of this framework are as follows. First, this framework is a coherent workflow of evaluation, interpretation, and verification of the street built environment features and visual walkability perception based on the big data of panoramic SVIs. It is a new paradigm and can be scaled to other visual-based subjective perceptions. Second, immersive VR panorama-based evaluation not only mitigates the gaps between browser-based auditing and on-site assessment in real-world, but also achieved more consistent evaluation results without scoring bias of the natural SVIs based on different views of the same location. Third, the correlation analysis and interpretable deep learning results not only provide both macro- and micro- interpretation of the relationship between VWP and street built environment features, but also help stakeholders to trust deep learning models with visualized results and promote human-machine collaboration.

This study demonstrated the feasibility and reliability of using VR techniques with panoramic SVIs and deep learning methods to measure the visual perception of walkability and explore the contributions of visual elements, revealing the potential for applying these methods practically. First, the results of the on-site visual walkability perception evaluation of the campus of the anonymous university, combined with the inferred mapping results of the trained VWPCL model, validated the

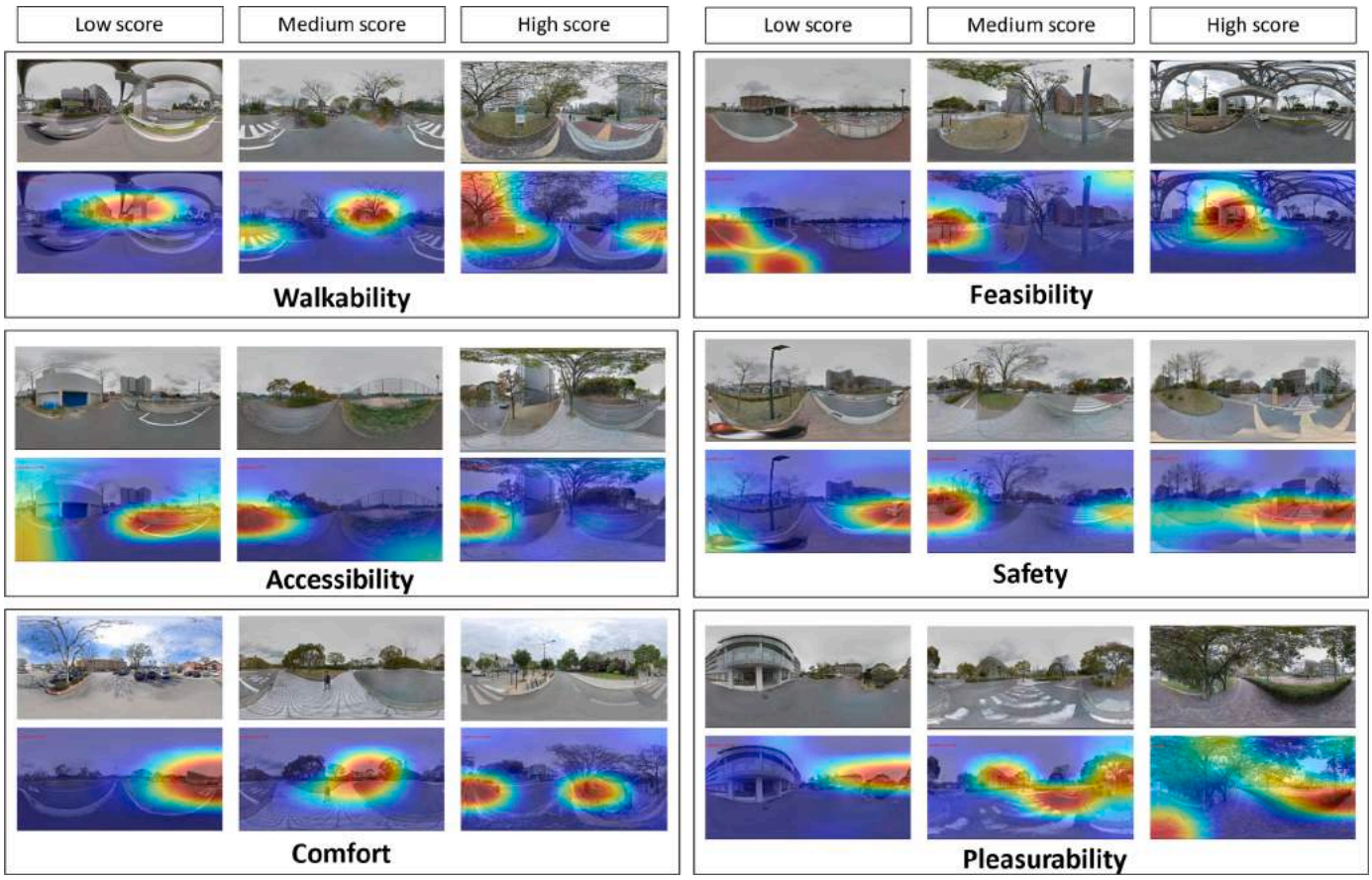


Fig. 16. Grad-CAM result samples for the on-site verification experimental area.

Table 6
Statistical results of 50 images for Grad-CAM verification.

VWP category	Contributing objects in the questionnaires and activated areas in the activation heat map		
	No. totally consistent	No. partially consistent	No. totally inconsistent
Walkability	31 (62%)	12 (24%)	7 (14%)
Feasibility	24 (48%)	20 (40%)	6 (12%)
Accessibility	27 (54%)	10 (20%)	13 (26%)
Safety	28 (56%)	11 (22%)	11 (22%)
Comfort	34 (68%)	8 (16%)	8 (16%)
Pleasurability	38 (76%)	7 (14%)	5 (10%)

accuracy of our proposed VWPCL model for VWP prediction. The evaluation and prediction results of the trained VWPCL model can be used in guidance services for recommending walking paths with high VWP scores. Second, the regression results and activation maps take researchers and urban planners a step further in understanding the interplay of perceived willingness to walk in a street built environment and street-level semantics and features. Collecting data on spatiotemporal SVIs and VWP changes, which can be combined with local socioeconomic, physical activity, and other data, will help urban planners understand potential patterns of homogeneity and heterogeneity in cities and reveal the impact of street attributes, such as what demographic and economic attributes of neighborhoods are more likely to be associated with pedestrian-friendly streets in the built environment. In addition, the Grad-CAM results facilitate understanding of which part of an SVI leads the CNN to make the final classification decision, which helps to locate specific targets in the image that affect specific VWP attributes. This also helps in debugging the decision-making process of the CNN and building classification datasets with higher accuracy, especially in the case of classification errors.

As described in Section 4.2, a stepwise regression analysis and deep learning interpretable analysis were conducted to identify the visual streetscape elements and VWP in terms of six categories (walkability, feasibility, accessibility, safety, comfort, and pleasurability). Most of the findings in Section 4.2 are consistent with those of previous studies (Alfonzo, 2005; Blečić et al., 2018; Ewing & Handy, 2009; Zhou et al., 2019). The contribution of identified visual elements, such as urban greenery, sidewalks, traffic lights, and traffic signs, to specific VWP indicators will directly support the theory and practice of designing street built environments. In addition, screening panoramic streetscape images of specific high-scoring VWPs and using them as a database to control the scale and characteristics of visual elements in data-driven generation of street built environments can further inform urban designers. For instance, techniques such as the generative adversarial network can be used to generate urban scenes that are considered to be highly visually walkable, safe, and comfortable.

Here, we discuss some interesting findings of the Grad-CAM results. First, although the physical components of streets belonging to the same category in semantic segmentation are all activated simultaneously in Grad-CAM every time, as shown in the high accessibility score results in Fig. 12, only the sidewalk on the right side with a larger relative width is activated. This indicates that the pixel share of physical components in panoramic SVIs as explanatory variables in the stepwise regression model is not exactly a criterion for the deep learning model to perform classification task learning, and sometimes feature learning of physical components was used to perform classification. Second, some questionnaire results about the contributing streetscape elements differ from the highlighted part of the activation diagram. The reasons may be that, for one, people may sometimes be influenced by their first impressions or certain attention-grabbing features when making perceptual

judgments about streetscape images, and may be unconscious of how other elements have an increased or decreased contribution to the overall perception, that is, the primacy effect (Jones et al., 1968) and halo effect (Leuthesser et al., 1995). Third, the DCNN classification model based on feature learning differs from the way people judge the perceptual classification of images in terms of learning. As shown in Fig. 17(a), the trees that played a positive role in the medium walkability score evaluation and the vehicles that played a negative role were activated at the same time, while one of the participants focused only on the large greenery area. Fig. 17(b) shows that the participants noticed only the garbage cans, which seriously affected the visual walking quality, while the computer considered that the proportion of harmonious greenery and building façades as positive physical components also contributed to the scoring segment.

Currently, in the study of the urban built environment with the aim of increasing walkability, it is crucial to systematically understand human perception of various urban built environment elements, reveal the intrinsic mechanism of perception of streetscape elements that contribute to visual walkability attractiveness, and establish a set of urban built environment planning and design methods adapted to public participation and a human-centered perspective. Panoramic VR technology can be used to improve the efficiency and accuracy of visual walkability perception for street-level urban environment analysis. Low-cost and mobile-based VR platforms, such as Google Cardboard and smartphones, allow for extension beyond the laboratory setting in order to enhance community participation in VWP evaluation and street design research. In addition, combined with the proposed deep learning approach, the proposed research framework can help designers to quickly conduct in-depth perceptual analysis and evaluation of VWP of urban streets and landscape nodes, and to accurately identify various urban landscape elements and their contribution to visual walkability. This is important for promoting the humanization and sophistication of the identification, evaluation, and management of urban built environment elements related to walkability.

The limitations of this research are listed below:

- The scope and number of SVI ratings of the constructed VRVWPR dataset are still relatively limited. The dataset can be further expanded to lay the foundation for a larger and more accurate quantitative analysis study of VWP.
- Differences in the results on the contribution of the physical components of streets between stepwise regression and Grad-CAM (especially streetscape elements other than some semantic segmentation elements, such as blind alleys and street furniture) illustrate that the existing semantic categories of semantic segmentation as explanatory variables for VWP are relatively one-sided in some cases. The explanatory strength of the stepwise regression model can be further improved by introducing a semantic segmentation dataset labeled with physical components related to the VWP audit.
- The deep learning model used in this study for panoramic SVI analysis is a traditional CNN based on 2D planar images. However, for panoramic SVIs, the distortion caused by simply applying the CNN to the planar projection of spherical images will inevitably cause bias in the recognition and analysis of image elements. For example, the percentage of pixels of physical elements near the corners of the panoramic SVIs increases. Cohen et al., (2018) proposed that a spherical CNN based on spherical cross-correlation that is both expressive and rotation-equivariant can be used for 3D images to better solve this problem. In future research, the spherical CNN should be used to establish a research framework for further improving the accuracy of the results. Similarly, in the correlation analysis, the element ratio of panoramic SVIs can be mapped to 3D through some mathematical transformation to obtain a more accurate element ratio in the future study.
- Although Grad-CAM provides a new perspective to assist in validating DCNN classification results for VWP and to assist in analyzing and discovering the role of streetscape elements in VWP, the current activation map can only provide coarse-grained interpretation results, which cannot meet the need for finer interpretation in detailed design scenarios of the street built environment. The guided Grad-CAM method, which can be used as an enhanced version of Grad-CAM, does not require modification of the network structure or retraining of the model, but can provide finer-grained interpretation results.
- In this study, the last convolutional layer of the DCNN model was extracted to obtain the activation map results when the VWPCL model was interpreted using Grad-CAM. However, many studies have pointed out that it is possible to obtain very different activation results using different DCNN models or extracting different convolutional layers (Du et al., 2019; Linardatos et al., 2020). This is an issue worth exploring in future research.

6. Conclusion

The VWP of a neighborhood affects the walking behavior and physical and mental health of its occupants. Several studies have used raw SVIs to measure the human perception of a place in large-scale urban areas. However, the difference between browser-based evaluation on natural SVIs and field auditing and the black-box working mechanism of deep learning makes it challenging to audit large amounts of SVI data closer to the actual visual walking experience, to accurately analyze the streetscape elements, and to understand and trust the

(a) A sample of medium score

(b) A sample of low score

	Contributed objects in SVIs	
	Sample (a)	Sample (b)
Grad-CAM Results	Road, car & tree	Building, tree & garbage cans
Questionnaire results of a participant	Tree	Garbage cans

(c) Comparison of contributed objects

Fig. 17. Grad-CAM result samples of VWP activation of walkability category: (a) A sample of medium score (confidence level: 1.0). (b) A sample of low score (confidence level: 0.68). (c) Comparison of contributing objects between Grad-CAM results and questionnaire results of a participant.

perceptual classification results from machine learning. In this study, we employed a quantitative VR panorama-based approach to measure human perceptions of the visual walkability of street built environments at scale in an automated, efficient, and accurate manner. We developed the VRVWPR dataset and the VWPCL model that can be used as a tool to evaluate and predict the VWP of new areas in six categories. Second, a multivariate stepwise linear regression analysis combined with a semantic segmentation algorithm is used to identify visual elements that strongly influence human VWP. In addition, an interpretable deep learning approach is used to understand the high-level information in the images and better understand human perception of the built environment. Finally, we validated the VWPCL model results and Grad-CAM model results based on-site auditing. We found that urban greenery and sidewalks always elicited a positive VWP perception in the street built environment, while trucks and motorcycles were a negative element, which is consistent with the literature in related fields. The stepwise regression results can reveal broad trends in visual elements and VWP effects, while the Grad-CAM results can reveal possible potentially important visual features that are difficult to reveal in the regression results.

The framework proposed in this study comprises observing, perceiving, auditing, and understanding the street built environment and predicting subjective perceptions based on the big data of panoramic SVIs as a new paradigm. The results of the study support the theory and practice of street walkability-oriented neighborhood design. This study also demonstrated the reliability of using VR panoramic SVIs and machine learning methods to understand the visual walkability value of how people perceive the physical environment of places, which can help researchers understand the impact of potential streetscape element features on VWP, and also provides a basis for humanizing and quantifying research on the built environment of streets with a view toward walkability and the construction of smart cities. Future work includes spherical CNN-based VWPCL model, eye-tracker-based method verification, and tailored training dataset for semantic segmentation models.

CRediT authorship contribution statement

Yunqin Li: Conceptualization, Methodology, Software, Validation, Writing – original draft. **Nobuyoshi Yabuki:** Conceptualization, Writing – review & editing, Supervision, Project administration. **Tomohiro Fukuda:** Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Abley, S., & Hill, E. (2005). Designing living streets—A guide to creating lively, walkable neighbourhoods. https://trid.trb.org/view/769621.

Aghaabbasi, M., Moeinaddini, M., Shah, M. Z., Asadi-Shekari, Z., & Kermani, M. A. (2018). Evaluating the capability of walkability audit tools for assessing sidewalks. *Sustainable Cities and Society, 37*, 475–484.

Alfonzo, M. A. (2005). To walk or not to walk? The hierarchy of walking needs. *Environment and Behavior, 37*(6), 808–836, 10/djkp88.

Arellana, J., Saltarín, M., Larrañaga, A. M., Alvarez, V., & Henao, C. A. (2020). Urban walkability considering pedestrians' perceptions of the built environment: A 10-year review and a case study in a medium-sized city in Latin America. *Transport Reviews, 40*(2), 183–203, 10/ghtvgp.

Ashihara, Y. (1983). The aesthetic townscape.

Bellazzi, A., Bellia, L., Chinazzo, G., Corbisiero, F., D'Agostino, P., Devitofrancesco, A., et al. (2022). Virtual reality for assessing visual quality and lighting perception: A systematic review. *Building and Environment, 209*, Article 108674, 10/gn4zkg.

Blečić, I., Cecchini, A., Trunfio, G. A., et al. (2018). Towards automatic assessment of perceived walkability. In O. Gervasi, B. Murgante, S. Misra, E. Stankova, C. M. Torre, A. M. A. C. Rocha, et al. (Eds.), *Computational science and its applications—ICCSA 2018* (pp. 351–365). Springer International Publishing, 10/gn8jcq.

Bosselmann, P., Macdonald, E., & Kronemeyer, T. (1999). Livable streets revisited. *Journal of the American Planning Association, 65*(2), 168–180, 10/ch8d9g.

Campisi, T., Ignaccolo, M., Inturri, G., Tesoriere, G., & Torrisi, V. (2021). Evaluation of walkability and mobility requirements of visually impaired people in urban spaces. *Research in Transportation Business & Management, 40*, Article 100592, 10/gm8k39.

Cerin, E., Leslie, E., Toit, L., du, Owen, N., & Frank, L. D. (2007). Destinations that matter: Associations with walking for transport. *Health & Place, 13*(3), 713–724, 10/djn947.

Chan, E. T., Schwanen, T., & Banister, D. (2021). Towards a multiple-scenario approach for walkability assessment: An empirical application in Shenzhen. *China. Sustainable Cities and Society, 71*, Article 102949.

Chen, L., Chen, J., Hajimirsadeghi, H., & Mori, G. (2020). Adapting Grad-CAM for embedding networks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2794–2803).

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

Cohen, T.S., .Geiger, M., Koehler, J., & Welling, M. (2018). Spherical CNNs. ArXiv: 1801.10130 [Cs, Stat]. http://arxiv.org/abs/1801.10130.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM, 63*(1), 68–77. https://doi.org/10.1145/3359786

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

Duncan, Dustin T., Aldstadt, Jared, Whalen, John, Melly, Steven J., & Gortmaker, Steven L. (2011). Validation of Walk Score® for estimating neighborhood walkability: An analysis of four US metropolitan areas. *International Journal of Environmental Research and Public Health, 8*(11), 4160–4179, 10/bmzxp3.

Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design, 14*(1), 65–84, 10/dhwh5h.

Fan, P., Wan, G., Xu, L., Park, H., Xie, Y., Liu, Y., et al. (2018). Walkability in urban landscapes: A comparative study of four large cities in China. *Landscape Ecology, 33*(2), 323–340, 10/gc22vf.

Frackelton, A., Grossman, A., Palinginis, E., Castrillon, F., Elango, V., & Guensler, R. (2013). Measuring walkability: Development of an automated sidewalk quality assessment tool. *Suburban Sustainability, 1*(1), 10/ggxjxk.

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., & Li, B. (2020). Axiom-based Grad-CAM: Towards accurate visualization and explanation of CNNs. *ArXiv Preprint ArXiv: 2008.02312*.

Gong, F.-. Y., Zeng, Z.-. C., Zhang, F., Li, X., Ng, E., & Norford, L. K. (2018). Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Building and Environment, 134*, 155–167, 10/gdh2jd.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018a). Local rule-based explanations of black box decision systems. *ArXiv Preprint ArXiv, 1805*, 10820.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018b). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR), 51*(5), 1–42. https://doi.org/10.1145/3236009

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

He, N., & Li, G. (2021). Urban neighbourhood environment assessment based on street view image processing: A review of research trends. *Environmental Challenges, 4*, Article 100090. https://doi.org/10.1016/j.envc.2021.100090

Hu, C. B., Zhang, F., Gong, F. Y., Ratti, C., & Li, X. (2020). Classification and mapping of urban canyon geometry using Google Street View images and deep multitask learning. *Building and Environment, 167*, Article 106424. https://doi.org/10.1016/j.buildenv.2019.106424

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 2261–2269). CVPR, 10/gfhw3n.

Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019). Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 279–287), 10/ghj6ft.

Jiang, B., Chang, C.-. Y., & Sullivan, W. C. (2014). A dose of nature: Tree cover, stress reduction, and gender differences. *Landscape and Urban Planning, 132*, 26–36, 10/f6qmzq.

Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology, 10*(4), 317. https://doi.org/10.1037/h0026818

Ki, D., & Lee, S. (2021). Analyzing the effects of Green View Index of neighborhood streets on walking time using Google Street View and deep learning. *Landscape and Urban Planning, 205*, Article 103920, 10/ghm9rn.

Kim, S.-. N., & Lee, H. (2022). Capturing reality: Validation of omnidirectional video-based immersive virtual reality as a streetscape quality auditing method. *Landscape and Urban Planning, 218*, Article 104290, 10/gpj5w5.

Lee, H., & Kim, S.-. N. (2021). Perceived safety and pedestrian performance in pedestrian priority streets (Ppss) in Seoul, Korea: A virtual reality experiment and trace

mapping. *International Journal of Environmental Research and Public Health, 18*(5), 1–16. Scopus10/gpj7q7.
Leuthesser, L., Kohli, C. S., & Harich, K. R. (1995). Brand equity: The halo effect measure. *European Journal of Marketing*. https://doi.org/10.1108/03090569510086657
Li, W., Zhai, J., & Zhu, M. (2022a). Characteristics and perception evaluation of the soundscapes of public spaces on both sides of the elevated road: A case study in Suzhou, China. *Sustainable Cities and Society, 84*, Article 103996. https://doi.org/10.1016/j.scs.2022.103996
Li, X., Zhang, C., & Li, W. (2015). Does the visibility of greenery increase perceived safety in urban areas? Evidence from the Place Pulse 1.0 dataset. *ISPRS International Journal of Geo-Information, 4*(3), 1166–1183. https://doi.org/10.3390/ijgi4031166
Li, Y., Yabuki, N., & Fukuda, T. (2022b). Exploring the association between street built environment and street vitality using deep learning methods. *Sustainable Cities and Society, 79*, Article 103656, 10/gn6gmp.
Li, Y., Yabuki, N., Fukuda, T., & Zhang, J. (2020). A big data evaluation of urban street walkability using deep learning and environmental sensors-a case study around Osaka University Suita campus. Volume 2, 319–328.
Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018
Ma, X., Ma, C., Wu, C., Xi, Y., Yang, R., Peng, N., et al. (2021). Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing. *Cities (London, England), 110*, Article 103086. https://doi.org/10.1016/j.cities.2020.103086
Min, W., Mei, S., Liu, L., Wang, Y., & Jiang, S. (2020). Multi-task deep relative attribute learning for visual urban perception. *IEEE Transactions on Image Processing, 29*, 657–669, 10/gjpx8q.
Mouratidis, K., & Hassan, R. (2020). Contemporary versus traditional styles in architecture and public space: A virtual reality study with 360-degree videos. *Cities (London, England), 97*, Article 102499, 10/gpjjmm.
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*, 887–894. XLIII-B3-202110/gphkms.
Ortega, E., MartÍN, B., LÓPez-Lambas, M. E., & Soria-Lara, J. A. (2021). Evaluating the impact of urban design scenarios on walking accessibility: The case of the Madrid 'Centro'district. *Sustainable Cities and Society, 74*, Article 103156.
Quercia, D., Schifanella, R., & Aiello, L. M. (2014). The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM conference on hypertext and social media* (pp. 116–125), 10/gfvsn2.
Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation, 29*(9), 2352–2449. https://doi.org/10.1162/neco_a_00990
Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering, 20*(5), 589–600. https://doi.org/10.1109/TKDE.2007.190734
Saadi, I., Aganze, R., Moeinaddini, M., Asadi-Shekari, Z., & Cools, M. (2022). A participatory assessment of perceived neighbourhood walkability in a small urban environment. *Sustainability, 14*(1), 206, 10/gpj7qc.
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, 10/f5bs55.
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).
Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *ArXiv Preprint ArXiv, 1605*(01713).
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv, 1409*, 1556.
Southworth, M. (2005). Designing the walkable city. *Journal of Urban Planning and Development, 131*(4), 246–257.
Tang, J., & Long, Y. (2019). Measuring visual quality of street space and its temporal variation: Methodology and its application in the Hutong area in Beijing. *Landscape and Urban Planning, 191*, Article 103436, 10/gf5bvk.
Verma, D., Jana, A., & Ramamritham, K. (2020). Predicting human perception of the urban environment in a spatiotemporal urban setting using locally acquired street view images and audio clips. *Building and Environment, 186*, Article 107340. https://doi.org/10.1016/j.buildenv.2020.107340
Wang, H., & Yang, Y. (2019). Neighbourhood walkability: A review and bibliometric analysis. *Cities (London, England), 93*, 43–61, 10/gg2fzm.
Yameqani, A. S., & Alesheikh, A. A. (2019). Predicting subjective measures of walkability index from objective measures using artificial neural networks. *Sustainable Cities and Society, 48*, Article 101560.
Yao, Y., Liang, Z., Yuan, Z., Liu, P., Bie, Y., Zhang, J., et al. (2019). A human-machine adversarial scoring framework for urban perception assessment using street-view images. *International Journal of Geographical Information Science, 33*(12), 2363–2384, 10/ggqrj6.
Yin, L., & Wang, Z. (2016). Measuring visual enclosure for street walkability: Using machine learning algorithms and Google Street View imagery. *Applied Geography, 76*, 147–153, 10/f88n47.
Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., et al. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning, 180*, 148–160, 10/gfp3zx.
Zhang, J., Fukuda, T., & Yabuki, N. (2021). Development of a city-scale approach for façade color measurement with building functional classification using deep learning and street view images. *ISPRS International Journal of Geo-Information, 10*(8), 551, 10/gm6hpj.
Zhang, R.-. X., & Zhang, L.-. M. (2021). Panoramic visual perception and identification of architectural cityscape elements in a virtual-reality environment. *Future Generation Computer Systems, 118*, 107–117, 10/gphcxh.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921–2929).
Zhou, H., He, S., Cai, Y., Wang, M., & Su, S. (2019). Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. *Sustainable Cities and Society, 50*, Article 101605, 10/gjsjcr.