Spatial Information Research (2025) 33:57
https://doi.org/10.1007/s41324-025-00656-7

ORIGINAL ARTICLE

Modeling and explaining perceived walkability in urban environments using street view images and explainable AI

Jaeyeon Choi[1] · **Youngok Kang**[1]

Received: 30 July 2025 / Revised: 28 October 2025 / Accepted: 30 October 2025 / Published online: 17 November 2025
© The Author(s) 2025

Abstract
Walking is a fundamental mode of urban mobility that promotes physical and mental well-being, mitigates environmental issues, and enhances community livability. As research increasingly emphasizes not only the physical but also the perceptual dimensions of walkability, understanding how pedestrians cognitively evaluate their surroundings has become a critical challenge. This study proposes a GeoAI-based framework for modeling and explaining perceived walkability in urban environments using street view imagery (SVI) and explainable AI techniques. A perceptual dataset derived from 171,118 pairwise comparisons was used to train the RSS-Swin model, enabling continuous estimation of walkability scores that reflect nuanced human judgments. Urban visual features were extracted from SVIs through semantic segmentation with SegFormer-B5, incorporating both object-level components and image brightness (HSV_V). To enhance interpretability, AutoML-based optimization and SHAP analysis were employed to identify and explain the key urban form features influencing perceived walkability. The proposed framework not only achieves high predictive accuracy but also provides transparent insights into the visual and environmental factors shaping pedestrian perception, offering practical implications for data-informed urban design and policy.

Keywords Perceived walkability · Street view imagery · Explainable AI · AutoML · SHAP (SHapley additive exPlanations) · Semantic segmentation · RSS-Swin

1 Introduction

Walking is the most fundamental mode of transportation and is closely tied to the daily lives of urban residents. Numerous studies highlighting the benefits of walking have demonstrated that walking activities promote both physical and mental health, help address traffic and environmental issues, and contribute to revitalizing local economies and strengthening community bonds among residents [1–3]. As the significance of walking has grown, the pedestrian environment is now recognized as a core component of the built environment in contemporary urban planning. To promote walkability, governments and local authorities across the globe are increasingly incorporating pedestrian infrastructure into urban planning agendas and mandating systematic assessments of walkability conditions.

Alongside this emphasis on the physical pedestrian environment, there is a growing interest in how pedestrians perceive and cognitively evaluate their surroundings. The study by Cao et al. (2006) suggests that cognitive perceptions may have a more direct impact on the actual usability of urban spaces and citizens' activities than physical conditions alone [4]. Accordingly, in walkability research, increasing attention has been paid to subjective, perception-based assessments—so-called cognitive evaluations. In line with this trend, a number of studies have sought to qualitatively assess walkability and to identify the urban built environment factors that influence such evaluations.

Recently, the application of street view images (SVI) and deep learning-based AI technologies has emerged as a promising approach for qualitative assessments of urban environments. With improved access to platforms like Google Street View and rapid advancements in deep learning for

✉ Youngok Kang
ykang@ewha.ac.kr

Jaeyeon Choi
jaeyeon9983@naver.com

[1] Department of Social Studies (Geography), Ewha Womans University, 52 Ewhayeodae- gil, Seodaemun-gu, Seoul 03760, South Korea

image analysis, new methodologies have been developed to conduct large-scale qualitative evaluations of urban areas [5–8]. SVIs, in particular, are highly suitable for this purpose because they offer a perspective that closely mimics the pedestrian's point of view, making them especially valuable for capturing the experiential qualities of the walking environment [9, 10]. Compared to traditional methods such as on-site surveys or questionnaires, the use of SVIs and deep learning allows for more cost- and time-efficient analyses, as well as finer-grained spatial assessments across entire urban areas.

To make cognitive evaluations of walkability useful for practical urban planning, it is not sufficient to merely present assessment scores; it is also necessary to identify which specific urban features contribute to those evaluations. This is particularly critical when deep learning models are used, as their black-box nature often limits interpretability. A single deep learning model alone may not provide sufficient insight into the factors driving evaluation scores, posing challenges for real-world applications in planning contexts. As a result, recent studies have begun to analyze the factors influencing such cognitive evaluations of pedestrian environments. These studies have employed computer vision techniques such as image classification, semantic segmentation, and object detection to quantify visual elements in SVIs, and have used machine learning models to explain how these elements influence perception-based scores [11–13]. However, this line of research is still in its early stages.

In addition, recent advances in machine learning have introduced the use of Automated Machine Learning (AutoML) techniques. AutoML automates the iterative processes involved in model development—such as model selection and hyperparameter tuning—based on dataset characteristics. While AutoML has been actively adopted for efficient and accurate model construction in various fields, it has yet to be applied to analyze the determinants of cognitive evaluations in pedestrian environments. Such an application is both necessary and timely.

Unlike previous studies, this research offers three key contributions to the growing field of GeoAI-based walkability analysis. First, while earlier approaches often relied solely on subjective surveys or opaque deep learning models, this study integrates a large perceptual dataset derived from 171,118 pairwise comparisons with an interpretable prediction model based on RSS-Swin. This enables the estimation of continuous walkability scores that reflect nuanced human judgments, overcoming the limitations of discrete classification and black-box predictions. Second, semantic segmentation using SegFormer-B5 is employed to extract fine-grained urban form features from street view images, incorporating both object-level visual elements and image brightness (HSV_V). Unlike previous studies that used coarse indicators such as greenery presence or road width, this approach provides a more detailed representation of the built environment's compositional influence on perceived walkability. Third, by integrating Explainable AI techniques—specifically AutoML-based optimization and SHAP interpretation—the framework not only achieves high predictive accuracy but also offers transparent explanations of how visual and environmental features shape pedestrian perception. This dual emphasis on performance and interpretability provides practical insights for data-driven and human-centered urban design.

The remainder of this paper is organized as follows. Section 2 describes the study area and data. Section 3 outlines the methodology, including model development and feature extraction. Section 4 presents the results and key findings. Section 5 concludes with implications and future directions.

2 Related work

2.1 Qualitative evaluation of pedestrian environments using SVIs and deep learning

With the increased accessibility of SVI and rapid advancements in deep learning technologies, numerous studies have been conducted to assess perceived urban environments using these tools. Two major methodological approaches have emerged in studies that use SVIs to conduct perception-based assessments of pedestrian environments. The first involves assigning absolute Likert-scale scores to individual images based on perceived attributes such as safety, attractiveness, or comfort [13–16]. To ensure consistency in scoring, participants are often trained in advance or experts are recruited for evaluation. However, this approach can lead to additional costs and may face challenges in collecting a sufficiently large volume of labeled data.

Alternatively, a growing body of research has adopted pairwise comparison methods [7, 17]. In this approach, participants are shown two randomly selected images and asked to choose the one that better represents a given perceptual attribute. This relative comparison method has been found to be more efficient and reliable than assigning absolute scores to individual images [18, 19], and is increasingly being applied in recent studies.

A prominent example of a dataset based on pairwise comparison for urban perception is Place Pulse [8], which includes SVIs from New York, Boston, and the Austrian cities of Linz and Salzburg, evaluated for attributes such as safety, uniqueness, and affluence. Dubey et al. (2016) expanded the dataset to include six perceptual attributes and a broader set of cities, resulting in the Place Pulse 2.0 dataset [5]. They proposed the Ranking Street Score-Convolutional

Neural Network (RSS-CNN), a deep learning model trained on pairwise comparisons. Since the release of this dataset and model, numerous follow-up studies have built upon Dubey et al.'s methodology to assess the cognitive perception of urban environments [20–23]. In the context of walkability, Kang et al. (2023) developed a pairwise comparison dataset specific to perceived pedestrian environments and introduced the Global_Patch_RSS-CNN model, which captures both global and localized visual features of the image [7].

These methodological advancements have opened new directions in the qualitative evaluation of urban environments. They help overcome the cost and time limitations of traditional field surveys and questionnaires, and enable detailed spatial analysis across large urban areas. However, due to the complexity of deep learning models, it remains difficult to interpret how specific visual or structural features influence the resulting evaluations. Therefore, complementary studies are needed to explain the key factors that shape these perception-based assessments.

2.2 Identifying the influencing factors of perceived pedestrian environments

To enhance the practical utility of perceived walkability assessments, several studies have examined the relationship between perceptual scores and physical streetscape features. These studies typically apply regression models—often based on machine learning—to explore how visual features extracted from SVIs contribute to qualitative assessment results.

Joglekar et al. (2020), for example, analyzed the elements associated with streetscapes perceived as beautiful by pedestrians [20]. Using PlaceNet for image classification, they categorized the images into four types: Architectural, Walkable, Landmark, and Natural. Their findings showed that scenes perceived as beautiful were most often associated with the "Walkable" category.

Other studies have focused more specifically on walkability. Nagata et al. (2020) collected walkability scores for Google Street View (GSV) images from individuals aged 65 and older, and applied multiple regression analysis using semantic segmentation results [11]. The analysis yielded an R^2 value of 0.51 and revealed that vegetation near roads positively influenced walkability perception, while visible sky had a negative impact. Lee et al. (2022) employed pairwise comparison datasets and walkability scores generated by the Global_Patch_RSS-CNN model. They used semantic segmentation results to identify influencing features, applying eight machine learning models, including multiple linear regression, random forest, and XGBoost. Support vector regression achieved the highest performance with an R^2 of 0.61 [12]. Their findings showed that roads and sidewalks were associated with positive walkability perceptions, while sky views were negatively associated. More recently, Huang et al. (2024) used SVIs and virtual reality (VR) technologies to evaluate four dimensions of walkability—usefulness, safety, comfort, and attractiveness—and performed regression analysis using physical features derived from semantic segmentation [13]. Their models yielded R^2 values ranging from 0.45 to 0.57, and the study found that wide pedestrian spaces, abundant greenery, and iconic landmarks positively contributed to perceived walkability. Despite these advancements, further research is needed to more comprehensively understand how built environment elements affect perceived pedestrian environments.

2.3 Applications of automl techniques

In studies that aim to interpret the results of deep learning-based qualitative assessments of urban environments, machine learning algorithms are frequently employed to explain which features influence model outputs. These studies typically involve selecting a suitable algorithm and manually tuning its hyperparameters to optimize performance—an essential but time-consuming step that often depends heavily on the researcher's expertise. Moreover, such manual approaches may fall short in capturing complex patterns or nonlinear relationships inherent in the data.

To address these limitations, recent studies have adopted AutoML techniques. AutoML streamlines the model development pipeline by automating processes such as data preprocessing, feature selection and extraction, model selection, and hyperparameter tuning. It systematizes these tasks into functions that can efficiently produce optimized models through iterative search and evaluation [24].

AutoML has been applied across a variety of tasks, including classification, time-series forecasting, and clustering. It is also gaining traction in regression tasks relevant to this study's goals. For instance, Nam et al. (2020) used H2O AutoML to identify terrain and geological factors associated with landslides [25], while Zhong et al. (2024) employed AutoGluon to model urban surface radiation temperatures [26]. Xue et al. (2024) utilized PyCaret to develop regression models for identifying factors influencing bicycle safety [27].

These studies demonstrate that AutoML can facilitate efficient, high-performance model development across multiple domains. As more powerful and flexible AutoML libraries become available, their potential for broader applications continues to grow. However, despite its advantages, AutoML has yet to be widely used in studies analyzing factors that influence deep learning–based assessments of urban environments. Thus, there is a clear need for research that

introduces an AutoML-driven methodology for explaining perceptual evaluations derived from SVI and deep learning models.

3 Methodology

The overall research framework is illustrated in Fig. 1. First, SVIs were crawled for the study area of Anyang, South Korea. Second, a pairwise comparison training dataset was constructed to predict perceived walkability scores. Third, the pairwise comparison data and corresponding SVIs were fed into a deep learning model to estimate perceived walkability scores. Fourth, semantic segmentation was applied to the SVIs to extract the proportion of visual elements representing the built environment. Fifth, using AutoML, a machine learning model was trained using the perceived walkability scores as the dependent variable and the proportions of segmented objects as independent variables. Finally, an analysis was conducted to identify the key built environment features influencing perceived walkability and to draw practical implications.

3.1 Study area and data collection

This study focuses on the city of Anyang, South Korea. Anyang is characterized by a mix of an old urban core centered around Anyang Station and a newer development area in Pyeongchon, part of the metropolitan new town initiative. The city is also actively implementing pedestrian infrastructure improvement and urban renewal projects [28, 29]. To collect the necessary SVIs for analysis, road network data based on the road name address system—provided by the Ministry of the Interior and Safety—was used to determine sampling locations [30]. SVI collection points were extracted at 30-meter intervals along road centerlines, and at each point, four directional SVIs—front, back, left, and right—were retrieved from the closest image location within a 50-meter radius.

To ensure that the visual assessment of the urban landscape reflects a wide range of spatial characteristics, additional SVIs were collected from Yeongdeungpo District in Seoul, which contains well-defined land uses such as commercial and industrial zones. A total of 45,420 SVIs were collected from 11,355 locations in Anyang, while 41,648

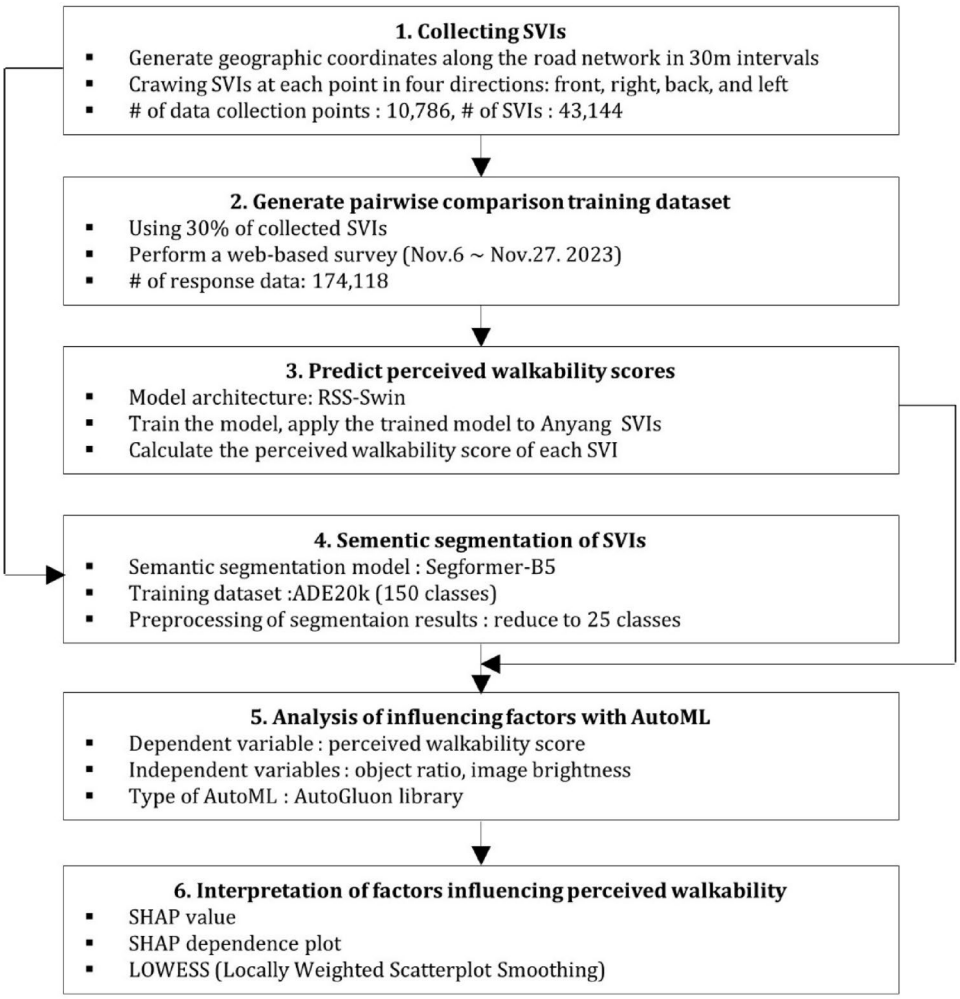


Fig. 1 Overall research framework

Table 1 Summary of data

Data	AnYang	Yeongdeungpo-gu	total
Sample points	11,355	10,412	21,767
Number of SVIs	45,420	41,648	87,068
Number of stratified points for training	3,107	2,371	5,478
Number of SVIs for training	12,428	9,484	21,912
Final number of SVIs for pairwise comparison	11,846	9,040	20,886

images were gathered from 10,412 locations in Yeongdeungpo, resulting in a combined dataset of 87,068 SVIs.

3.2 Cognitive evaluation scores for pedestrian environments

3.2.1 Training dataset construction

To train a deep learning model capable of predicting perceived walkability score, a training dataset comprising SVIs and corresponding human evaluation labels is required. The training dataset was constructed using 30% of the total collected SVIs. Sampling locations for training data were selected using stratified sampling, excluding highways, and accounting for road types (boulevard, street, alley) and land use categories (industrial areas, commercial/business zones, high-rise residential areas, low-rise residential areas, and others).

From Anyang, 12,428 SVIs were selected from 3107 locations, while 9484 images were chosen from 2371 locations in Yeongdeungpo District. After excluding images that were blurred or distorted, a total of 20,886 SVIs were finalized for use in the training dataset (Table 1).

While preferences can be measured using absolute scoring methods—where participants assign numeric ratings to individual SVIs—collecting responses through pairwise comparisons has been shown to be a more efficient and accurate approach [18, 19]. In this study, we adopted a relative comparison method, in which participants were shown two images at a time and asked to select the one they perceived as more walkable.

To facilitate this process, a web-based survey platform was developed, allowing participants to evaluate image pairs online (Fig. 2). For ethical compliance and protection of participants' personal information, the study received Institutional Review Board (IRB) approval from Ewha Womans University (October 17, 2023). Participants were recruited with consideration for age and gender diversity: 25 participants in their 20 s, 20 in their 30–40 s, and 20 aged 50 or older, resulting in a total of 65 evaluators.

To ensure response reliability, each participant was presented with five control questions within each set of 50 questions, which included pairs of images with clearly distinguishable walkability levels. These served as attention checks to detect careless or malicious responses and maintain data quality. Additionally, the question set was algorithmically structured to achieve statistical stability in preference scores even with a limited number of comparisons, minimizing variance caused by respondent heterogeneity [31]. The survey was conducted from November 6 to November 27, 2023. After excluding responses labeled as "equal walkability (=)," a total of 174,118 valid pairwise comparison responses were collected. On average, each image was compared more than 16 times across participants, which we consider sufficient to ensure stable estimation of walkability scores during model training.

3.2.2 Deep learning model training and prediction of cognitive pedestrian environment scores

The use of pairwise comparison data for cognitively evaluating the built environment has been widely facilitated by the release of the Place Pulse 2.0 dataset and the accompanying RSS-CNN model [5]. RSS-CNN consists of a Siamese

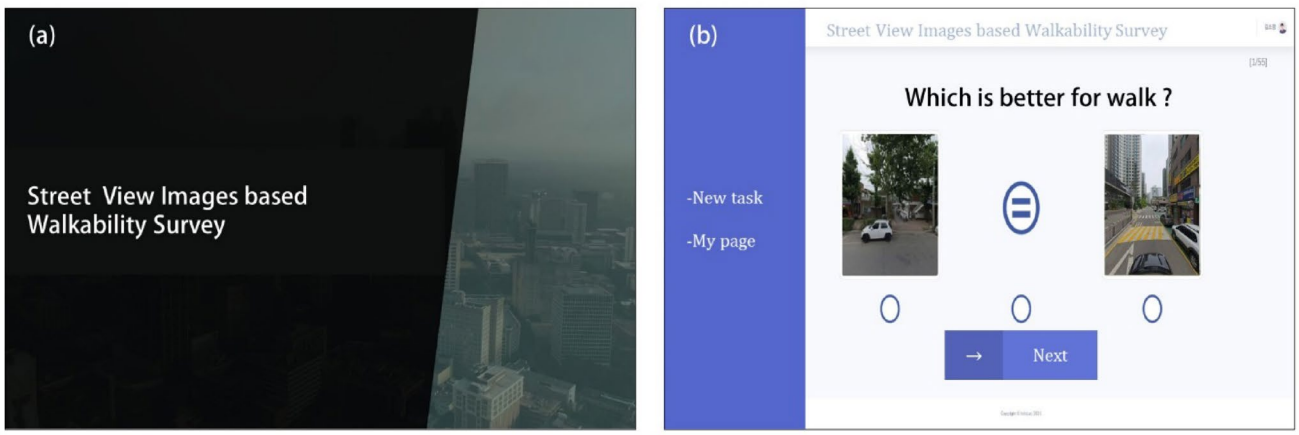


Fig. 2 Survey website: **a** landing page, **b** response screen

network—which learns the similarities and differences between two input images by jointly updating the weights of two identical neural networks [32]—and RankNet, which learns a ranking function based on pairwise comparisons [33].

Building on this, Kang et al. (2023) proposed the Global_Patch_RSS-CNN model, which improved upon the original RSS-CNN by incorporating both global and local visual features through the use of image patches [7]. Both models are convolutional neural networks (CNNs) that use VGG16 as their backbone. More recently, Kim et al. (2025) introduced the RSS-Swin model, which replaces the CNN backbone with a transformer-based architecture. This model extends the self-attention mechanism through local and shifted windows, enabling it to simultaneously capture local and global features. Its hierarchical structure has demonstrated strong performance even with high-resolution images [34].

Accordingly, the present study adopted the RSS-Swin model for perceived walkability score prediction. The architecture of the RSS-Swin model used in this study is illustrated in Fig. 3. It is composed of a Siamese network and RankNet framework, with the Swin Transformer serving as the backbone for image feature extraction. The model consists of two main components: a feature block that learns image representations and a score block that predicts pairwise preference scores using the RankNet al.gorithm. Model training was conducted on Amazon Web Services (AWS) using a g4dn.12xlarge instance with the following specifications: Ubuntu 18.04 OS, PyTorch 1.13.1, four NVIDIA T4 GPUs, 48 vCPUs, and a 900 GB NVMe SSD. The trained RSS-Swin model achieved a classification accuracy of 79.71%, an improvement over the 76.77% accuracy obtained using the Global_Patch_RSS-CNN model trained on the same dataset [35].

Using the trained RSS-Swin model, perceived walkability scores were predicted for 43,144 SVIs in Anyang City corresponding to boulevard, street and alley. The scores ranged from −5.67 to 2.92, with a mean of −.33 and a standard deviation of approximately 1.36. A spatial visualization of the predicted scores is presented in Fig. 4. For each location, the final score was calculated as the average of the walkability scores assigned to the four directional SVIs (front, back, left, and right).

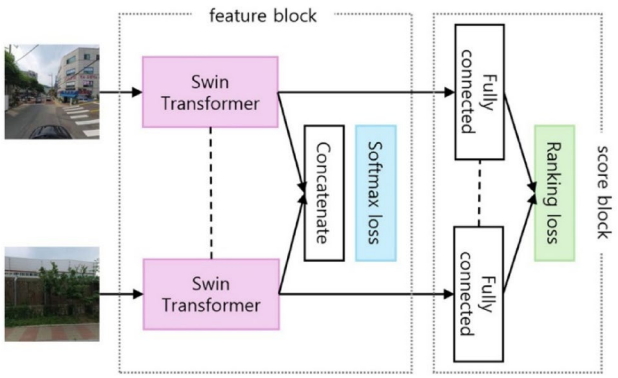


Fig. 3 RSS-Swin model architecture [34]

3.3 Feature extraction from SVI

To analyze the influence of urban built environment characteristics on perceived walkability, visual features were extracted from SVIs, specifically the proportion of streetscape objects and overall image brightness. These features were used as independent variables in subsequent analyses.

To identify and quantify the presence of various object classes within the SVIs, semantic segmentation techniques were employed. Semantic segmentation of SVI typically utilizes datasets such as Cityscapes [36] and ADE20K [37]. The Cityscapes dataset includes 30 object classes commonly found in urban street scenes, whereas the ADE20K dataset encompasses 150 object categories across both indoor and outdoor environments. Given this study's focus on outdoor objects such as vehicles, buildings, and natural elements, the ADE20K dataset was considered more appropriate due to its broader and more detailed classification of outdoor scene elements. Therefore, a semantic segmentation model trained on the ADE20K dataset was adopted for feature extraction.

For semantic segmentation, the SegFormer-B5 model [38] was selected. SegFormer integrates a transformer-based encoder with a lightweight decoder, delivering high performance in image segmentation tasks. Its encoder architecture, known as the Mix Transformer, combines the patch-based structure of the Vision Transformer [39] with the hierarchical representation of CNN, thereby enabling more effective image learning through a broader effective receptive field compared to traditional CNN-based models.

In addition, SegFormer uses a simple multi-layer perceptron as its decoder, significantly reducing computational complexity while maintaining strong segmentation performance. Among models trained on the ADE20K dataset, SegFormer-B5 demonstrated a favorable trade-off between model size and performance, achieving higher mean Intersection over Union (mIoU) scores than other widely used models such as DeepLabV3, Swin Transformer, and SETR. Specifically, SegFormer-B5 achieved an mIoU of 51.8, surpassing DeepLabV3 (44.1) and SETR (50.2) on the ADE20K dataset.

Furthermore, we compared the segmentation results of SegFormer-B5 and DeepLabV3 using SVIs from Anyang. As shown in Fig. 5, SegFormer-B5 yielded fewer segmentation errors, confirming its superiority for this application. Based on these findings, SegFormer-B5 was adopted in this

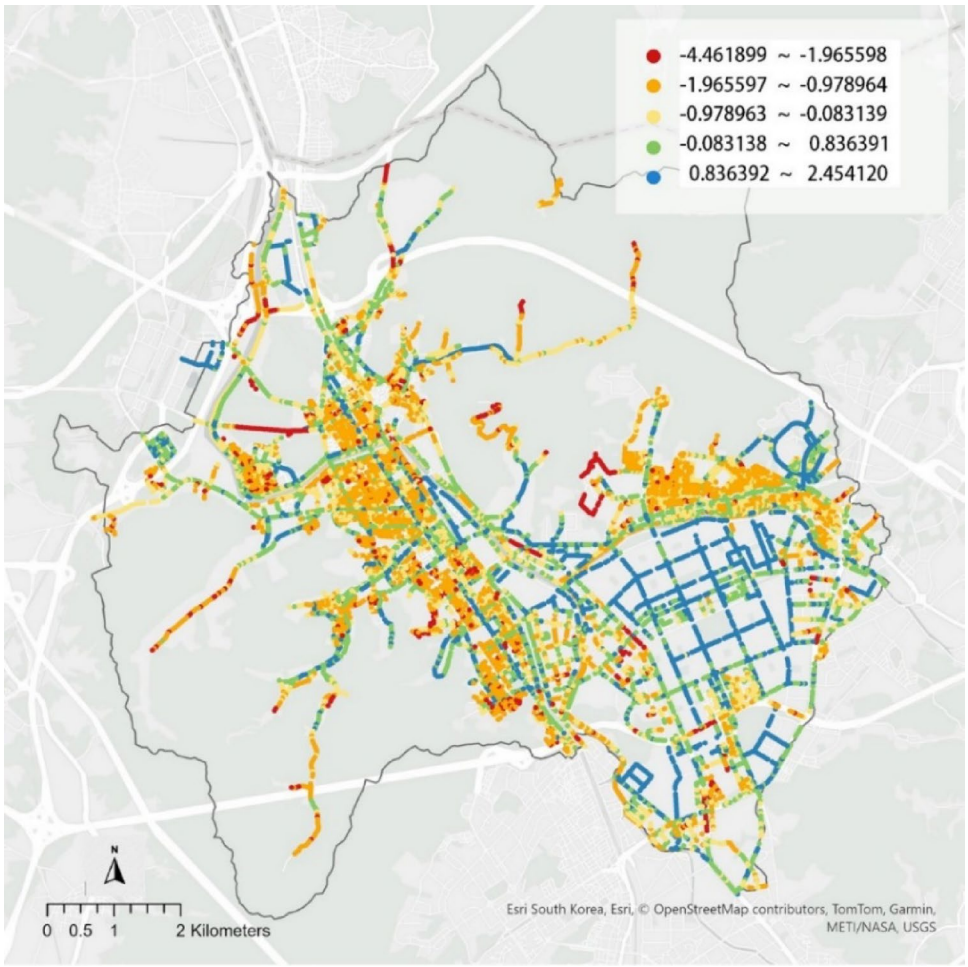


Fig. 4 Visualization of perceived walkability score at street level

study to accurately extract a diverse range of visual elements from SVIs.

Among the semantic segmentation results, it was necessary to exclude indoor object classes included in the ADE20K dataset, as well as elements not commonly observed in Korean streetscapes. To this end, any class for which over 90% of its pixel proportions across all images were zero was excluded from further analysis. As a result, 24 object classes were retained for use in the study.

To improve the accuracy of the regression model and prevent misinterpretation of results, multicollinearity among independent variables was examined using the Variance Inflation Factor (VIF). VIF quantifies how much the variance of a regression coefficient is inflated due to linear dependence among predictors. In general, a VIF value exceeding 5 is considered indicative of problematic multicollinearity in explanatory models. The VIF for the "road" class was 5.4911, suggesting multicollinearity, while the VIF for "sky" was also relatively high at 4.8398.

SVIs with a high proportion of road pixels typically correspond to wide roads such as boulevards, whereas those with a low proportion often depict narrow alleys. Thus, road area can be interpreted as a proxy for the openness of the streetscape. Similarly, the presence of sky is also commonly associated with perceived openness. Given that both object classes are conceptually linked to spatial openness, we merged them into a single composite variable labeled "openness" to address the multicollinearity issue. Following this adjustment, the VIFs were recalculated for the revised set of 23 object classes, confirming that all variables had VIF values below 3, indicating that multicollinearity was no longer a concern (Table 2).

In addition to the object proportions derived through semantic segmentation, image brightness was also considered as a potential influencing factor in the qualitative evaluation of SVIs. Brightness is a visually perceptible element that can be extracted directly from images and may vary significantly depending on contextual factors such as the presence of tunnels or overpasses, degree of sky openness, and the proportion of surrounding buildings or vegetation. As such, it was deemed to have a direct impact on perceptual assessments.

To extract brightness values, various color spaces can be utilized, including RGB (Red, Green, Blue), HSL (Hue,

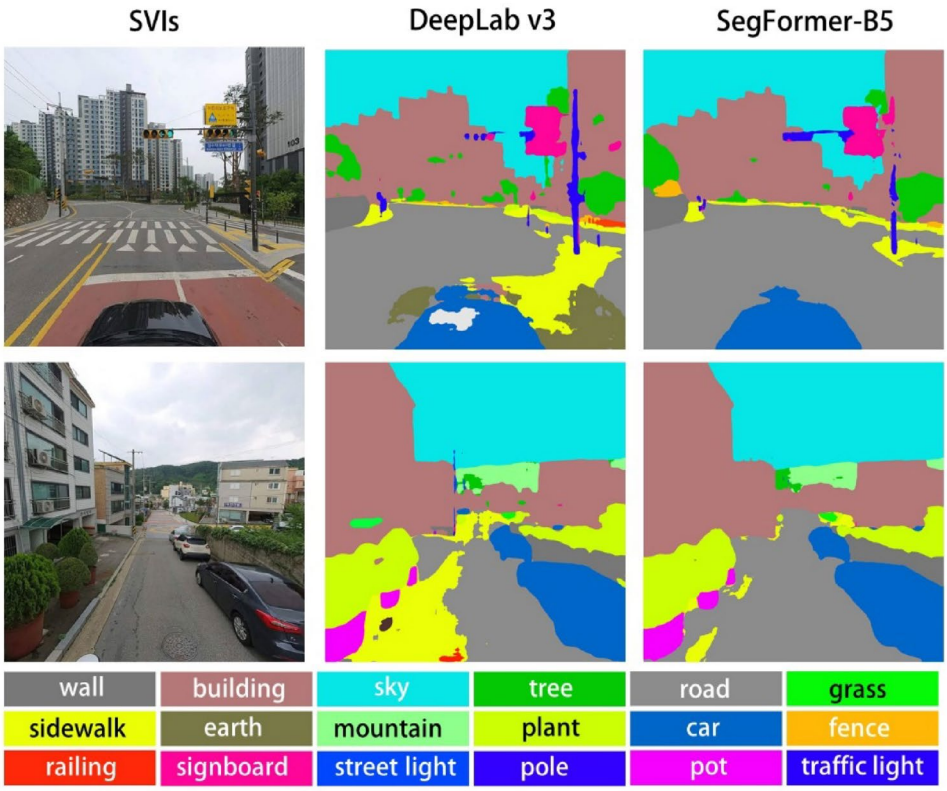


Fig. 5 Comparison of semantic segmentation results of DeepLabV3 and SegFormer-B5 models for SVIs

Table 2 VIF for the adjusted 23 object classes

Object	VIF
wall	1.067030
building	2.451618
tree	2.105525
grass	1.084746
sidewalk	1.661604
person	1.079813
earth	1.175178
mountain	1.038811
plant	1.232642
car	1.782423
fence	1.147526
railing	1.061348
box	1.076479
signboard	1.209937
stairs	1.039463
truck	1.031700
awning	1.115233
streetlight	1.182298
pole	1.134262
trade name	1.135318
traffic light	1.044392
ashcan	1.066931
Openness (sky + road)	2.180890

Saturation, Lightness), and HSV (Hue, Saturation, Value). In this study, the V (value) channel of the HSV color space—representing brightness—was used to quantify the brightness of each image. Since images are typically stored in the RGB color space, we employed the OpenCV library to convert each image to HSV format and then extracted the brightness values. These values are expressed as floating-point numbers ranging from 0 to 255. Including this brightness measure, a total of 24 independent variables were ultimately used in the analysis.

3.4 Analysis of influencing factors using expainable AI

Building upon the visual and brightness features extracted from the SVIs in the previous section, this stage aimed to identify how these factors influence perceived walkability. To achieve both high predictive accuracy and transparent interpretation, an Explainable AI (XAI) approach was adopted. Specifically, automated machine learning was first applied to efficiently construct and optimize predictive models, followed by post-hoc interpretability analysis to uncover the underlying relationships between visual variables and perceptual outcomes. Specifically, AutoGluon, an open-source AutoML framework developed by Amazon Web Services, was adopted to automate model training, selection, and hyperparameter tuning. The Tabular Predictor module

was utilized to evaluate multiple algorithms—including gradient boosting, random forests, and neural networks—and to select the optimal model or ensemble that minimized prediction error while avoiding overfitting. Internal k-fold cross-validation (default = 5) was performed automatically to ensure robust and unbiased evaluation.

The dependent and independent variables used in the analysis are summarized in Table 3. A set of models was generated to identify the factors affecting the dependent variable, and the optimal model was selected through comparative evaluation. Model selection was based on the following performance metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2).

$$MAE = \frac{\sum |y - \widehat{y}|}{n} \tag{1}$$

$$MSE = \frac{\sum (y - \widehat{y})^2}{n} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum (y - \widehat{y})^2}{n}} \tag{3}$$

Table 3 Descriptive statistics of variables

Category	Name	Min	Max	Mean	SD
Dependent variables	walkability	−7.04	3.85	−0.38	1.51
Independent variables	wall	2.57	7.88	100	0
	building	26.47	22.41	100	0
	tree	12.62	15.01	100	0
	grass	1.03	4.67	66.23	0
	sidewalk	3.97	5.77	42.12	0
	person	0.08	0.36	11.57	0
	earth	1.19	4.92	71.05	0
	mountain	0.09	0.71	67.35	0
	plant	2.26	5.85	86.68	0
	car	5.85	6.84	43.01	0
	fence	0.98	2.61	58.48	0
	railing	0.24	1.22	54.54	0
	box	0.09	0.62	29.81	0
	signboard	0.62	1.62	49.07	0
	stairs	0.09	0.66	35.00	0
	truck	0.35	2.08	65.24	0
	awning	0.13	0.89	34.57	0
	streetlight	0.07	0.18	6.40	0
	pole	0.29	0.95	18.16	0
	trade name	0.09	0.53	15.11	0
	traffic light	0.05	0.31	11.77	0
	ashcan	0.05	0.28	10.42	0
	sky + road	38.47	21.20	100	0
	hsv_v	122.07	27.47	248.47	5.43

$$R^2 = \frac{\sum_{i=1}^{n} \left(\widehat{y_i} - \bar{y} \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} \tag{4}$$

$y =$ actual value, $\bar{y} =$ mean value, $\widehat{y} =$ predicted value, $n = \#$ of data

To interpret the contribution of each variable to the model output, the SHAP (SHapley Additive exPlanations) method [40] was applied. Based on cooperative game theory and the concept of Shapley values [41], SHAP quantifies how much each feature contributes to an individual prediction. This method provides a transparent, model-agnostic explanation of the learned relationships, allowing the identification not only of the most influential features but also of the direction and nature of their effects on perceived walkability. Since SHAP is typically applied to a single model rather than an ensemble, the SHAP analysis was conducted using the best-performing individual model identified in the AutoGluon pipeline. The SHAP value φ i (f, x) for each independent variable i with respect to an individual prediction f(x) is computed as shown in Eq. (9):

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{5}$$

$\boldsymbol{\phi_i}$ SHAP value of independent variable i.

$\boldsymbol{f}$ machine learning model.

$\boldsymbol{x}$ input data.

$\boldsymbol{M}$ total number of features (24).

$\boldsymbol{z'}$ a binary vector indicating feature inclusion ($z'_i = 1$ indicates that feature *i* is included).

4 Results and interpretation of influencing factors

4.1 Model performance and key predictors of perceived walkability

The results of the influencing factor analysis for perceived walkability score, obtained through AutoGluon, are presented in Table 4. The best-performing model was the ensemble model *WeightedEnsemble_L2*, which integrates five base learners available within AutoGluon: LightGBMXT, CatBoost, NeuralNetFastAI, NeuralNetTorch, and LightGBMLarge. The model achieved the highest R^2 value

Table 4 Comparison of model performance. Sorted in R^2 order

Model	MAE	MSE	RMSE	R^2
WeightedEnsemble_L2	0.500280	0.429150	0.655095	0.773672
LightGBMXT	0.505384	0.437404	0.661365	0.769319
CatBoost	0.506172	0.437946	0.661775	0.769033
LightGBMLarge	0.508470	0.444374	0.66614	0.765643
LightGBM	0.512377	0.450767	0.671392	0.762271
XGBoost	0.520957	0.464233	0.681346	0.755170
RandomForestMSE	0.5295+08	0.487675	0.698337	0.742807
ExtraTreesMSE	0.531099	0.483972	0.695681	0.744760
NeuralNetFastAI	0.531769	0.483357	0.695239	0.745084
NeuralNEtTorch	0.545689	0.525290	0.724769	0.722969
KNeighborsDist	0.630076	0.691006	0.831268	0.635573
KNeighborsUnif	0.633655	0.697212	0.834992	0.632300

(0.7737), demonstrating strong explanatory power, and showed a low RMSE (0.6551), corresponding to 36.8% of the standard deviation, indicating small prediction errors and stable performance.

Among the individual models, LightGBMXT showed the strongest performance. LightGBMXT is an extended version of LightGBM—a high-performance gradient boosting decision tree model—that incorporates more diverse tree structures and probabilistic feature sampling strategies. These enhancements allow it to learn complex data patterns effectively and improve generalization. The model achieved an RMSE of 0.6614 and an R^2 of 0.7693.

To identify the key factors influencing perceived walkability, feature importance was calculated based on the best-performing model, *WeightedEnsemble_L2* (Fig. 6). Feature importance was measured using permutation importance, which quantifies the decrease in prediction performance when each variable is randomly shuffled.

The results showed that the most influential variable in predicting cognitive walkability was the *sidewalk*. The importance score for sidewalk was approximately 0.3954, indicating that the proportion of sidewalks in a SVI is the most sensitive factor in determining walkability perception. In addition to sidewalks, *openness* (a composite variable combining road and sky) and *trees* also emerged as major contributors to perceived walkability.

4.2 Explainable AI–based interpretation of feature effects

To further examine the importance and directional effects of each variable, a SHAP analysis was conducted using the LightGBMXT model. Figure 7 presents a SHAP summary plot for the entire dataset. This plot visualizes the distribution of SHAP values for each independent variable, indicating how each feature contributes to the model's prediction of perceived walkability score. The features are ranked in descending order of importance. In the plot, the color represents the relative pixel proportion of that object in the SVI—red indicates higher proportions, while blue indicates lower ones. Along the x-axis, SHAP values to the right of center indicate a positive contribution to perceived walkability, whereas values to the left indicate a negative contribution. The analysis revealed that *sidewalk*, *tree*, *openness*, and *hsv_v* (image brightness) were positively correlated with perceived walkability. As the proportion of these features increased, their SHAP values shifted in the positive direction, indicating that these elements enhance the perception of a walkable environment. In contrast, features such as *wall*, *car*, *plant*, and *fence* showed negative associations. Higher proportions of these elements led to more negative SHAP values (indicated by the dominance of red points on the left side), suggesting a detrimental effect on perceived walkability. However, in the case of features such as *building*, *grass*, and *earth*, the SHAP values exhibited both positive and negative directions depending on their proportion, implying more complex, context-dependent relationships.

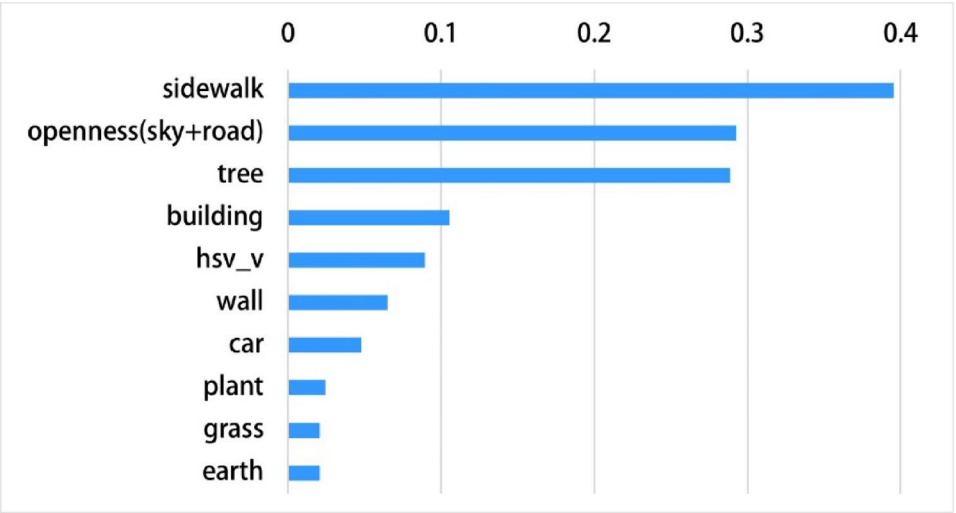


Fig. 6 The 10 major features of WeightedEnsemble_L2 model and the importance of each feature

This indicates the need for more detailed analysis of their effects on perceived walkability.

To further investigate features with ambiguous directional effects in Fig. 7, we compared SHAP summary plots for the top 10% and bottom 10% of images based on cognitive walkability scores (Fig. 8). In both groups, *sidewalk*, *tree*, *openness*, and *hsv_v* consistently showed positive associations with perceived walkability. Similarly, higher proportions of *wall* and *car* were negatively correlated with walkability in both groups. However, the degree of influence differed across groups. The *car* variable had a more pronounced negative impact in the top 10% group, suggesting that the presence of vehicles is particularly detrimental in otherwise walking environments. Conversely, *wall* exhibited greater importance in the bottom 10% group, indicating that its enclosed and obstructive nature strongly contributes to perceptions of poor walkability in such settings. In the case of *building*, the feature showed no clear directional effect in the bottom group, but in the top group, higher proportions were associated with more favorable walkability perceptions. For *grass*, a notable contrast was observed: it had a positive influence in the top group but was negatively associated with walkability in the bottom group. Upon reviewing the corresponding SVIs containing *grass* objects (Fig. 9), we found that in the top group, grass typically appeared as well-maintained landscaping between sidewalks and buildings. In contrast, in the bottom group, grass often represented unmanaged areas such as hillsides or riverbanks. This distinction in grass type likely influenced its perceived impact on walkability in the model.

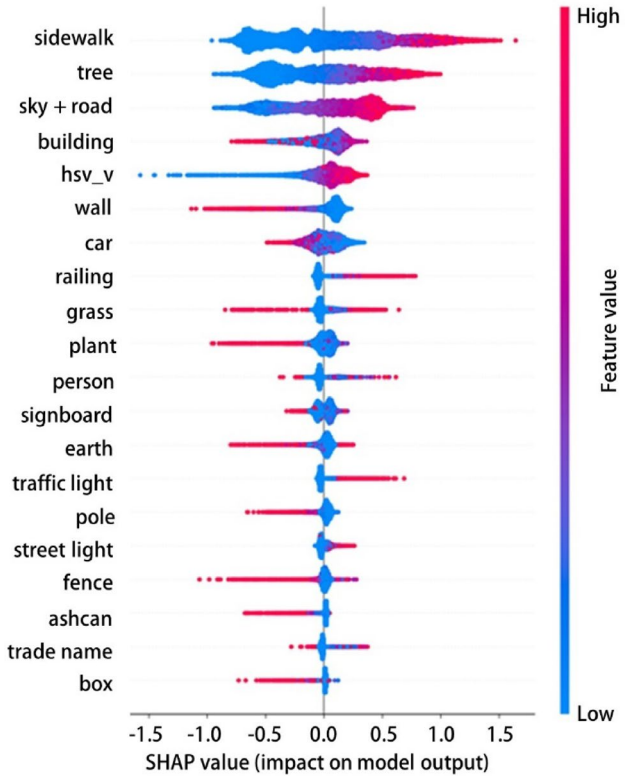


Fig. 7 SHAP summary plot of total data

To further investigate the relationship between key influencing factors and their SHAP values, dependence plots were visualized (Fig. 10). To better capture the trends within these plots, we applied LOWESS (Locally Weighted Scatterplot Smoothing) curves [42]. LOWESS is a non-parametric regression technique that assigns weights to data points within local windows, making it particularly effective for visualizing nonlinear relationships and complex patterns.

The results showed that *sidewalk*, *tree*, and *openness* exhibited a clear trend: as their proportions increased, perceived walkability scores also increased. Likewise, higher *hsv_v* values were associated with greater perceived walkability. Notably, extremely low brightness levels were linked to sharp declines in walkability scores. In contrast, the proportion of *wall* objects demonstrated a strong negative relationship with walkability. Except for images with minimal wall presence, an increasing wall ratio consistently led to lower perceived walkability scores. The *grass* variable also displayed a similar pattern: when its proportion was below 20%, it contributed positively to walkability perception, but proportions exceeding 20% were associated with decreased walkability score. This reflects the earlier observation that small, well-maintained patches of grass (e.g., planters) enhance perceived quality, whereas large, unmanaged grassy areas (e.g., riverbanks or hillsides) have the opposite effect.

For *building*, the relationship was more complex. When the building proportion was below approximately 40%, increases in this feature correlated positively with walkability. However, beyond that threshold, the relationship reversed, and higher building proportions were perceived as detrimental. This may be due to the dual nature of buildings in streetscapes: while they can signal spatial activation and vibrancy, they may also contribute to a sense of enclosure or visual obstruction. The dependence plot for *car* showed a nonlinear pattern. In general, SHAP values for this feature were below zero, indicating a negative effect on perceived walkability. However, in the range below 10%, a few cases showed a weak positive contribution. This anomaly is likely due to the nature of the SVI acquisition process. Since the SVIs used in this study were captured from vehicles, rear-facing images occasionally included the survey vehicle itself. These cases typically had a car pixel proportion of around 7%. Therefore, when additional car objects were present in the image beyond the survey vehicle, the environment was generally perceived as less walkable.

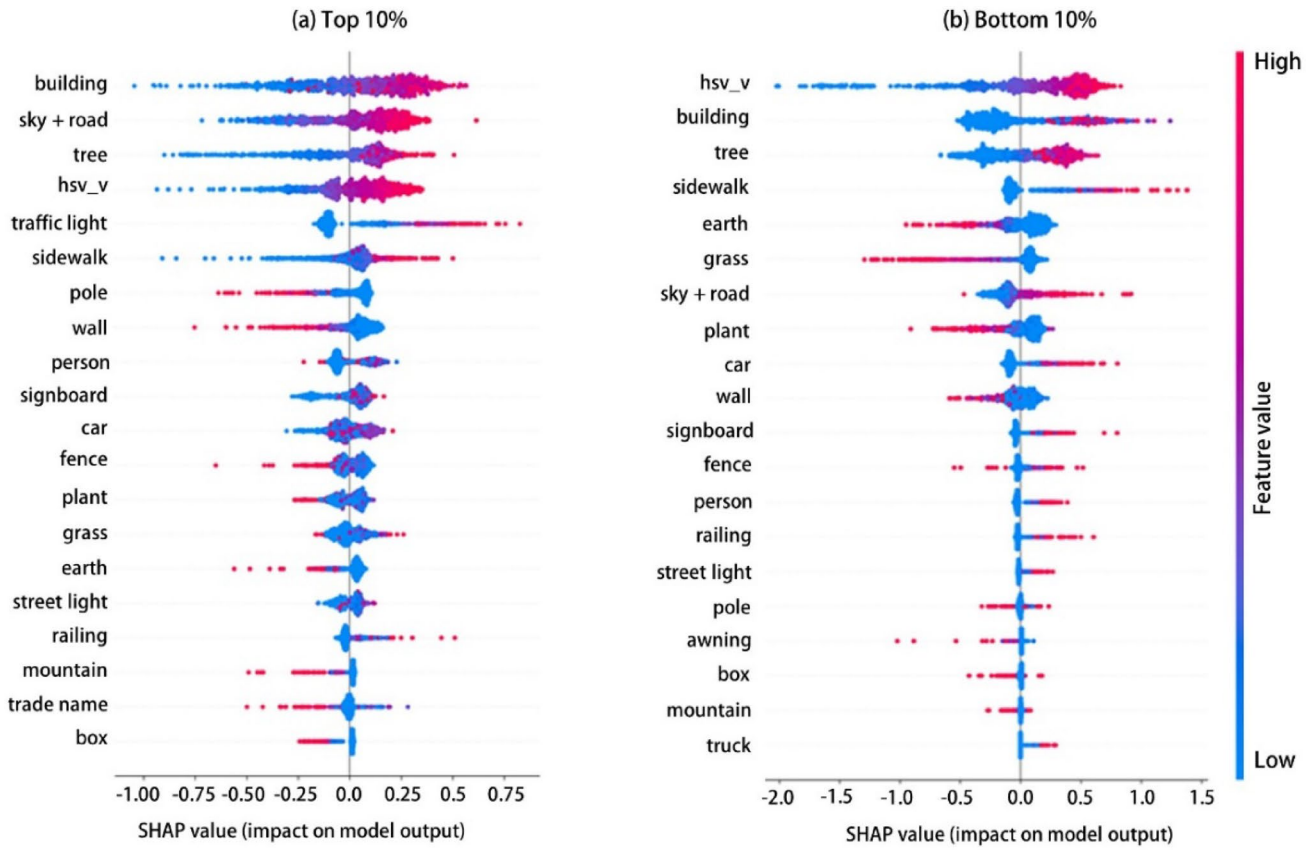


Fig. 8 SHAP summary plots of top 10% data and bottom 10% data and their sampled SVIs

5 Conclusion

This study developed an Explainable AI–based GeoAI framework to model and interpret perceived walkability in urban environments using SVI. By addressing the limitations of conventional field surveys and the black-box nature of deep learning models, the proposed approach provides a scalable and interpretable method for understanding how urban visual elements shape pedestrian perception.

A large-scale SVI dataset was collected from Anyang and Yeongdeungpo, South Korea, and perceptual walkability scores were generated through pairwise comparison experiments. The RSS-Swin model was employed to predict continuous perception scores, while visual features were extracted via the SegFormer-B5 semantic segmentation model trained on ADE20K, supplemented with image brightness (HSV_V). Using AutoGluon as an optimization engine within the XAI framework, the best-performing model (WeightedEnsemble_L2) achieved strong predictive accuracy ($R^2 = 0.7737$). SHAP analysis identified sidewalk, openness, trees, and brightness as the most positive contributors to perceived walkability, whereas wall, car, and fence negatively affected pedestrian perception. Dependence plots further revealed nonlinear and context-dependent effects for building, grass, and car, highlighting the complex and situational nature of perceived walkability.

This research contributes to the growing GeoAI literature by integrating explainable modeling and visual interpretation to bridge the gap between data-driven prediction and human-centered understanding. Unlike previous studies that relied solely on subjective surveys or opaque deep learning outputs, our framework demonstrates how interpretable GeoAI can support transparent and regionally contextualized decision-making in urban planning. In particular, it underscores the value of locally calibrated, culturally grounded perception data in understanding the unique spatial characteristics of dense Asian cities such as those in Korea.

While this study focused on two Korean cities, the findings may not fully capture the diversity of global urban environments. Cultural and morphological differences can influence how pedestrians perceive walkability; therefore, the results should be interpreted as regionally contextualized insights rather than universally generalizable conclusions. Future research should expand this framework to include cross-cultural comparisons across different cities, examining how spatial and cultural contexts jointly shape walkability perception.

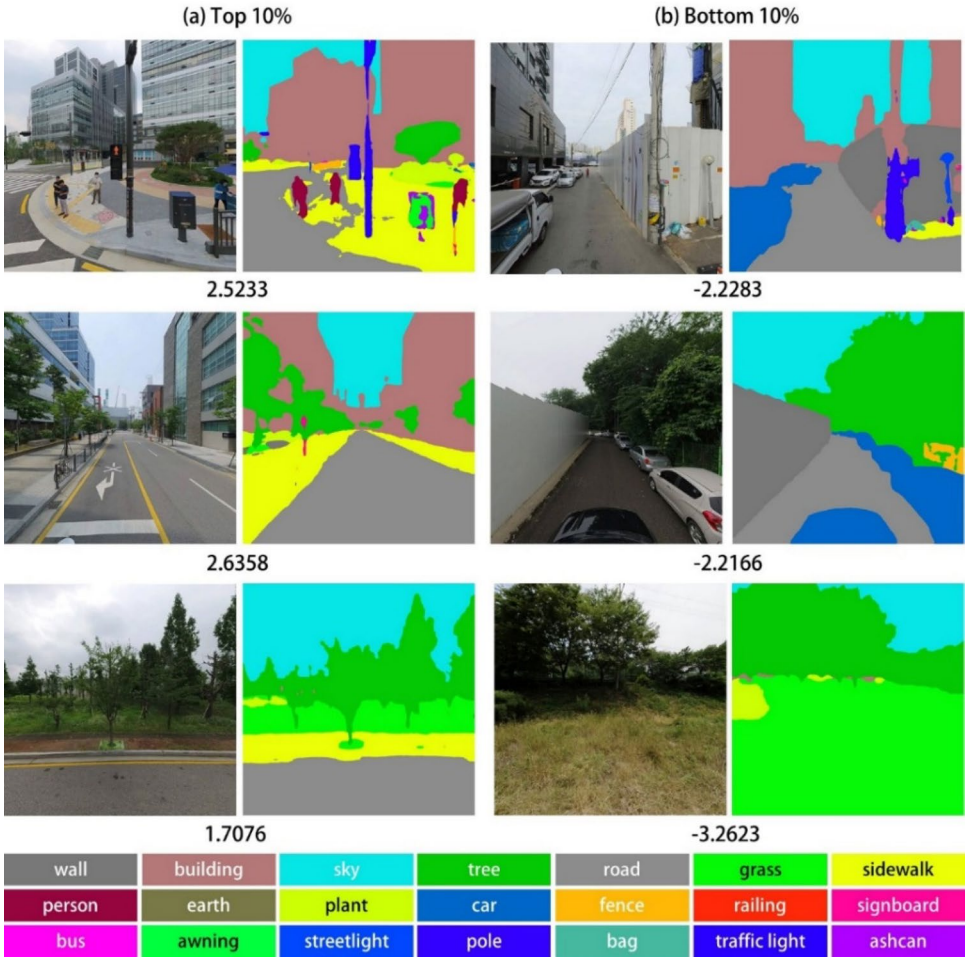


Fig. 9 Examples of semantic segmentation of top 10% data and bottom 10% data

In addition, future studies should incorporate temporal dynamics to capture variations in perception between day and night environments. Integrating nighttime SVIs, lighting conditions, and perceived safety would provide a more comprehensive understanding of walkability across temporal contexts. Furthermore, linking perception-based evaluations with empirical mobility data—such as pedestrian counts or trajectory-based movement patterns—could enhance the practical utility of walkability maps for prioritizing planning interventions. Such an integrated and explainable GeoAI approach would advance citizen-centered and data-informed urban management, supporting the creation of more equitable and livable city environments.

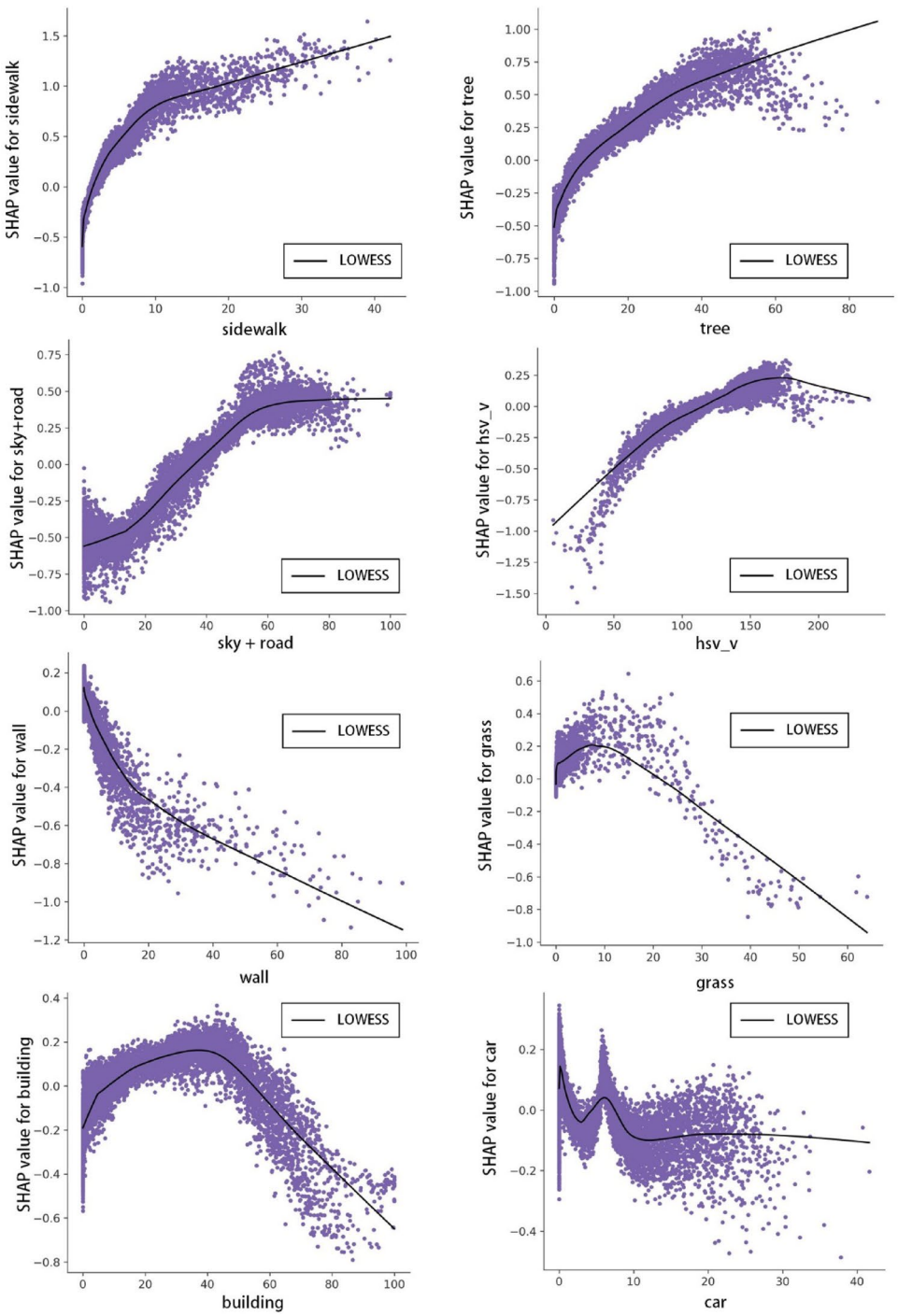


Fig. 10 SHAP value dependence plot of 8 factors

Funding This Research was supported by The Korea Agency for Infrastructure Technology Advancement (KAIA) funded by Ministry of Land, Infrastructure and Transport of Korean government (Grant No. RS-2022-00143782).

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Duncan, D. T., Aldstadt, J., Whalen, J., Melly, S. J., & Gortmaker, S. 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.
2. Keall, M. D., Shaw, C., Chapman, R., & Howden-Chapman, P. (2018). Reductions in carbon dioxide emissions from an intervention to promote cycling and walking: A case study from New Zealand. *Transportation Research Part D: Transport and Environment, 65*, 687–696.
3. Fonseca, F., Papageorgiou, G., Tondelli, S., Ribeiro, P., Conticelli, E., Jabbari, M., & Ramos, R. (2022). Perceived walkability and respective urban determinants: Insights from Bologna and Porto. *Sustainability, 14*(15), Article 9089.
4. Cao, X., Handy, S. L., & Mokhtarian, P. L. (2006). The influences of the built environment and residential self-selection on pedestrian behavior: Evidence from Austin, TX. *Transportation, 33*, 1–20.
5. Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In *Proceedings of Computer Vision – 14th European Conference* (pp. 196–212). Amsterdam, The Netherlands.
6. Guan, W., Chen, Z., Feng, F., Liu, W., & Nie, L. (2021). Urban perception: Sensing cities via a deep interactive multi-task learning framework. *Acm Transactions on Multimedia Computing Communications and Applications, 17*(1s), 1–20.
7. Kang, Y., Kim, J., Park, J., & Lee, J. (2023). Assessment of perceived and physical walkability using street view images and deep learning technology. *ISPRS International Journal of Geo-Information, 12*(5), Article 186.
8. Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PLoS One, 8*(7), Article e68400.
9. Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning, 215*, Article 104217.
10. Kang, Y. (2023). GeoAI application areas and research trends. *Journal of the Korean Geographical Society*, *58*(4), 395–418.
11. Nagata, S., Nakaya, T., Hanibuchi, T., Amagasa, S., Kikuchi, H., & Inoue, S. (2020). Objective scoring of streetscape walkability related to leisure walking: Statistical modeling approach with semantic segmentation of Google street view images. *Health & Place*, *66*, 102428.
12. Lee, J., Kang, Y., Kim, J., & Park, J. (2022). Analysis of street view image object affecting perceived walkability using machine learning. *Journal of the Association of Korean Geographers, 11*(3), 375–391.
13. Huang, G., Yu, Y., Lyu, M., Sun, D., Dewancker, B., & Gao, W. (2024). Impact of physical features on visual walkability perception in urban commercial streets by using street-view images and deep learning. *Buildings, 15*(1), Article 113.
14. Blečić, I., Cecchini, A., & Trunfio, G. A. (2018). Towards automatic assessment of perceived walkability. *International Conference on Computational Science and Its Applications*, (pp. 351–365).
15. Wang, R., Liu, Y., Lu, Y., Yuan, Y., Zhang, J., Liu, P., & Yao, Y. (2019). The linkage between the perception of neighbourhood and physical activity in Guangzhou, China: Using street view imagery with deep learning techniques. *International Journal of Health Geographics, 18*, 1–11.
16. Fang, C., Homma, R., Qiu, T., & Zhong, Q. (2024). The mediating role of visual behavior in perceived walkability evaluation. *Journal of Asian Architecture and Building Engineering, 24*, 1–15. https://doi.org/10.1080/13467581.2024.2402770
17. Li, Y., Yabuki, N., & Fukuda, T. (2022). Measuring visual walkability perception using panoramic street view images, virtual reality, and deep learning. *Sustainable Cities and Society, 86*, Article 104140.
18. Bijmolt, T. H., & Wedel, M. (1995). The effects of alternative methods of collecting similarity data for multidimensional scaling. *International Journal of Research in Marketing, 12*(4), 363–371.
19. Stewart, N., Brown, G. D., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*(4), 881–911.
20. Joglekar, S., Quercia, D., Redi, M., Aiello, L. M., Kauer, T., & Sastry, N. (2020). Facelift: A transparent deep learning framework to beautify urban scenes. *Royal Society Open Science, 7*(1), Article 190987.
21. Min, W., Mei, S., Liu, L., Wang, Y., & Jinag, S. (2020). Multitask deep relative attribute learning for visual urban perception. *IEEE Transactions on Image Processing*, *29*, 657–669.
22. Rossetti, T., Lobel, H., Rocco, V., & Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. *Landscape and Urban Planning, 181*, 169–178.
23. 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.
24. Nagarajah, T., & Poravi, G. (2019). A review on automated machine learning (AutoML) systems. In *2019 IEEE 5th International Conference for Convergence in Technology*, (pp. 1–6). IEEE.
25. Nam, K., Kim, M., Kwon, O., Wang, F., & Jeong, G. (2020). Prediction of landslides and determination of its variable importance using automl. *Journal of Engineering Geology, 30*(3), 315–325.
26. Zhong, X., Zhao, L., Ren, P., Zhang, X., & Wang, J. (2024). A physics-guided automated machine learning approach for obtaining surface radiometric temperatures on sunny days based on UAV-derived images. *Computers, Environment and Urban Systems, 113*, Article 102175.
27. Xue, H., Guo, P., Li, Y., & Ma, J. (2024). Integrating visual factors in crash rate analysis at intersections: An automl and SHAP approach towards cycling safety. *Accident Analysis & Prevention*, *200*, 107544.
28. Anyang City (2017). 2030 Anyang City Basic Plan. https://www.anyang.go.kr/main/contents.do?key=427
29. Anyang City (2023). 2040 Anyang City Basic Plan. https://www.anyang.go.kr/main/contents.do?key=427
30. Ministry of Interior and Safety Road segment of Road Name Address, VWorld. https://www.vworld.kr/dtmk/dtmk_ntads_s002.do?dsId=30055
31. Yoo, G., Lee, D., Lee, C., & Nam, K. (2022). Generating pairwise comparison set for crowd-sourcing based deep learning. *J Korea Ind Inf Syst Res*, *27*, 1–11. https://doi.org/10.9723/jksiis.2022.27.5.001
32. Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning Deep Learning Workshop*, 2(1), 1–30. Lille, France.
33. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent, In *Proceedings of the 22nd International Conference on Machine Learning*, (pp. 89–96). Bonn, Germany.
34. Kim, S., & Kang, Y. (2025). Predicting Fear of Crime and Evaluating CPTED-based Crime Safety Design Using Street View

Images and Deep Learning. *Journal of Korean Geographical Society*, 60(5), In press.
35. Choi, J., Kim, S., Noh, S., & Kang, Y. (2024). Comparison of physical and perceived walking environment using street view images and deep learning technology. *Journal of the Korean Cartographic Association, 24*(3), 45–60.
36. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 3213–3223).
37. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 633–641).
38. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems, 34*, 12077–12090.
39. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, T., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: transformers for image recognition at scale. In *9th International Conference on Learning Representations*, 186703. Virtual, Online.
40. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*. https://doi.org/10.48550/arXiv.1705.07874
41. Shapley, L. (1953). A value for N-Person games. *Contributions To the Theory of Games*, *2*, 307–317.
42. Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*(368), 829–836.

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