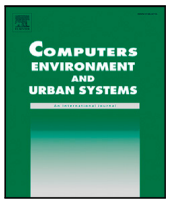
Computers, Environment and Urban Systems 123 (2026) 102356

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



Decoding human safety perception with eye-tracking systems, street view images, and explainable AI

Yuhao Kang [a,b], Junda Chen [c], Liu Liu [b], Kshitij Sharma [d], Martina Mazzarello [b], Simone Mora [b,d], Fábio Duarte [b,*], Carlo Ratti [b,e]

[a] GISense Lab, Department of Geography and the Environment, The University of Texas at Austin, Austin, TX, United States
[b] Senseable City Lab, Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, United States
[c] Department of Computer Science and Engineering, University of California San Diego, San Diego, CA, United States
[d] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway
[e] Politecnico di Milano, Milan, Italy

ARTICLE INFO

Keywords:
Eye-tracking systems
Perceptions
Street view images
Explainable artificial intelligence

ABSTRACT

The way residents perceive safety plays an important role in how they use public spaces, and it informs city planning and public policy. Recent studies have combined street view images and advanced computer vision techniques to measure human safety perceptions of urban environments. Despite their success, such studies have often overlooked the specific environmental visual factors that draw human attention and trigger people's feelings of safety perceptions. In this study, we introduce a computational framework that enriches the existing body of literature on place perception by using eye-tracking systems with street view images and explainable AI approaches. Eye-tracking systems measure what users are looking at and how long they engage with specific environmental elements. This allows us to explore the nuance of which visual environmental factors influence human safety perceptions. We conducted our research and recruited volunteers in Helsingborg, Sweden. By examining participants' focus on specific features using Mean Object Ratio in Highlighted Regions (MoRH) and Mean Object Hue (MoH), we identified key visual elements that attract human attention when perceiving safe environments. For instance, certain urban infrastructure (e.g., stairways and signboards) and public space (flags and chairs) features draw more human attention while the sky is less relevant in influencing safety perceptions. These insights offer a more human-centered understanding of which urban features influence human safety perceptions. Furthermore, we compared the real human attention from eye-tracking systems with attention maps obtained from eXplainable Artificial Intelligence (XAI) results. Several XAI models were tested, and we observed that XGradCAM and EigenCAM most closely align with human safety perceptual patterns. Our framework provides a valuable approach to enhance the interpretability and trustworthiness of XAI models by comparing them with empirically observed human behavior data. This study demonstrates the limitations of previous place perception studies that solely rely on street view images and computer vision techniques, which may not comprehensively capture the nuanced human experiences and behaviors at place. The inclusion of technologies such as eye-tracking not only deepens our comprehension of human subjective experiences, but also contributes to the development of safer environments and communities.

1. Introduction

Understanding human safety perceptions in the built environment offers key perspectives for creating safe places, and fostering a sense of security and well-being among residents (Ceccato, 2013; Raco, 2007; Un-Habitat, 2012). In alignment with the United Nations' Sustainable Development Goals (SDG), particularly the aim of fostering safe and inclusive cities and settlements (Abubakar & Aina, 2019), it is crucial to explore which places people perceive as safe and unsafe, and what urban features influence their perception beyond addressing criminal activities (Brymer, Crabtree, & King, 2021; Ceccato & Newton, 2015; Gobster & Westphal, 2004; Kang, 2023; Rahm, Sternudd, & Johansson, 2021; Rodriguez-Spahia & Barberet, 2020; Tabrizian, Baran, Smith, & Meentemeyer, 2018). Historical perspectives in this field have yielded a variety of theories that illuminate the nature of the safe-environment nexus. The *Broken Windows* theory, for instance, suggests that visible signs of disorder may lead to increased fear of crime (Doran & Lees,

* Corresponding author.
E-mail address: fduarte@mit.edu (F. Duarte).

https://doi.org/10.1016/j.compenvurbsys.2025.102356
Received 31 March 2025; Received in revised form 4 September 2025; Accepted 29 September 2025
Available online 1 November 2025

2005; O'Brien, Farrell, & Welsh, 2019; Wilson James & Kelling George, 1982). Similarly, the concept of *Defensible Space and Crime Prevention Through Environmental Design* emphasizes the significance of architectural and environmental design in deterring crime and promoting a sense of security (Jeffery, 1971; Newman, 1973). Additionally, the *Prospect-Refuge* theory suggests that environments offering clear visibility and places to hide or retreat are perceived as safer, influencing human behaviors and experiences choices (Appleton, 1975; Ramanujam, 2006). Consequently, an investigation of the subjective human sense of safety could provide novel insights into urban place-making and ultimately inform policies, foster safer communities and enhancing residents' sense of belonging.

To measure human place perceptions, recent studies have combined street view images and Artificial Intelligence (AI), particularly deep learning-based computer vision approach to measure human safety perceptions (Hamim & Ukkusuri, 2024; Ito, Kang, Zhang, Zhang, & Biljecki, 2024; Kang, 2023; Kang et al., 2023; Ramírez, Hurtubia, Lobel, & Rossetti, 2021; Wang et al., 2019; Zhang, Fan, Kang, Hu, & Ratti, 2021; Zhou et al., 2025). This approach utilizes large volumes of street view images coupled with advanced computer vision techniques to examine the complex relationships between human safety perceptions and their environmental contexts. The underlying hypothesis is that individuals' subjective responses to these street view images can accurately represent their perceptions of real-world environments. Notably, this approach offers a more efficient and cost-effective alternative to traditional methods such as surveys and interviews (Biljecki & Ito, 2021; Kang, Zhang, Gao, Lin, & Liu, 2020). However, despite its success, this approach faces two potential limitations in capturing the relationships between human safety perceptions and the built environment (Ito et al., 2024).

First, when analyzing the associations between human safety perceptions and objects within street view images, a notable issue refers to the undifferentiated treatment of all detected objects. Researchers tend to model the safety-environment nexus merely by considering the percentage of the presence of objects in these images (Dong et al., 2023; Larkin, Gu, Chen, & Hystad, 2021; Ogawa, Oki, Zhao, Sekimoto, & Shimizu, 2024; Ramírez et al., 2021; Zhang, Zhou et al., 2018). However, such approaches may not accurately reflect true human behaviors, and may instead resemble mathematical manipulation rather than performing behavioral analysis. This method may overlook the fact that, when observing these images, individuals often concentrate their attention selectively on certain objects rather than perceiving all objects as equally important (He, Qin, Shi, & Dong, 2024; Uttley, Simpson, & Qasem, 2018). Some smaller but unique objects in urban environments may draw more attention than larger, more commonplace ones. Furthermore, even when two images have similar proportions of environmental objects, they may evoke significantly different impressions due to their diverse environmental settings, which influences human perceptions. A larger proportion of urban elements in images does not necessarily indicate their significance in shaping human perceptual or behavioral responses. Human experiments remain essential to capture and reflect how individuals interact with their surrounding environments. Thus, such a discrepancy highlights the critical need to enrich prior studies with a more nuanced, human-centered perspective on environmental perceptions (Kang, 2025).

Second, a research gap arises from the "black box" nature of deep learning in measuring human place perception studies. Recently, the ethical implications of using artificial intelligence (AI) have attracted increasing attention, highlighting the need to critically assess the trustworthiness and potential biases in AI-generated outcomes (Jobin, Ienca, & Vayena, 2019; Kang, Gao, & Roth, 2024; Wach et al., 2023). In prior studies that utilized street view images and computer vision techniques, there have been efforts employing eXplainable Artificial Intelligence (XAI) methods such as Grad-CAM (Selvaraju et al., 2017) to delineate factors contributing to safety perceptions (Li, Yabuki, & Fukuda, 2022; Moreno-Vera, 2021; Sangers, van Gemert, & van Cranenburgh, 2022). The hypothesis is that these XAI methods can accurately reflect human behaviors and perceptions. Existing studies suggested that environmental features like greenery, cars, and sidewalks are important in influencing urban perceptions (Sangers et al., 2022). Despite these advancements, disparities highlighted by Li et al. (2022) between the outcomes of XAI-based methods and actual human perceptual preferences raise questions regarding the reliability of these metrics: To what degree do these measures accurately reflect human perceptions, and can we trust these metrics in understanding subjective human experiences? Consequently, a deeper exploration of the associations between human subjective safety perceptions and tangible elements in urban environments is essential. It is crucial to have a better understanding of the XAI methods utilized in prior studies by integrating more human perspectives.

To bridge the aforementioned two research gaps, we propose a computational framework that incorporates eye-tracking systems to understand how visual environmental factors trigger human safety perceptions. Eye-tracking systems, an emerging technology, have been used in a variety of spatial cognition and environmental psychology studies, offering unique opportunities to capture human visual attention and perceptions (Dong, Liao, Roth, & Wang, 2014; Hollander et al., 2018; Kiefer, Giannopoulos, Raubal, & Duchowski, 2017). The emergence of eye-tracking systems allows us to precisely track where and how long individuals focus their gaze when viewing images or environments (Goldberg & Kotval, 1999; He et al., 2023). Inspired by prior studies that leverage psychological methods to understand human perceptions (Qin, Dong, & Huang, 2023; Yang et al., 2024), we integrate eye-tracking systems to enrich our understanding of place perceptions. In our research, the use of eye-tracking systems could help identify specific features that attract an individual's attention. Such an approach is crucial to align more closely with real-world human behaviors and reflect human mental space which could more accurately identify and assess elements that influence safety perceptions (Hei, Yang, Dong, He, & Han, 2025; Shaw & Sui, 2021; Yang et al., 2025). Furthermore, eye-tracking data, which more accurately reflect human true behaviors, might be utilized to validate the explainability of deep learning models, providing a potential alternative to unlocking the "black box" of these models. We aim to validate the XAI-generated (eXplainable Artificial Intelligence) results by aligning them with eye-tracking results to help evaluate the robustness and trustworthiness of the workflow that combines street view images and computer vision techniques in measuring safety perceptions. Consequently, integrating eye-tracking systems into place perception studies presents a promising avenue that will not only fill the existing research gaps but also gain a comprehensive understanding of the characteristics and reliability of methods in place perception studies.

To this end, this study aims to deepen our understanding of human safety perceptions by using emerging datasets and technologies including eye-tracking systems, street view images, and XAI. We aim to investigate the influence of environmental features in street view images on human safety perceptions. Specifically, we ask the following research questions:

(1) What environmental features do individuals mostly focus on when assessing safety perception in cities?

(2) To what extent can eye-tracking system-based gaze heatmaps validate the outputs of XAI methods to enhance the understanding of their interpretability and trustworthiness for modeling safety perceptions?

To answer these questions, we first surveyed participants in Helsingborg, Sweden, to evaluate their perception of safety based on street view images of this city. Participants were equipped with eye-tracking systems to evaluate their safety perceptions of street view images, allowing us to further record and analyze their gaze patterns. After that, we generated a series of heatmaps to reflect their gaze patterns and identified environmental objects that drew significant attention. Two metrics were developed, the Mean Object Ratio in Highlighted Regions

(MoRH) and the Mean Object Hue (MoH) to detect important urban features. We further compared the heatmaps from the eye-tracking systems and those generated by XAI methods such as Grad-CAM to understand their similarities and discrepancies. Finally, we provided crucial insights for future place-making and urban design that contribute to the creation of safer environments, and discussed potential model biases to inform future technological advancements in the field.

The major contributions and innovations of this study are twofolds: First, by integrating multiple emerging technologies such as eye-tracking systems, street view images, and XAI, this study provides a deep understanding of how environmental factors shape safety perceptions at places; We identified specific objects in urban landscapes that may trigger human safety perceptions and delineated how individuals visually engage with their surroundings. This analysis provides practical guidance for environmental and urban design to enhance the sense of safety in communities. Second, our study validates the results obtained from XAI with real-world eye-tracking data on safety perceptions and identified the XAI model that most closely aligns with human perceptions; By doing so, our study offers implications for enhancing the trustworthiness of XAI models and advocates for enriching human-centered perspectives in the development of ethical AI methods to inform future urban planning and geography.

2. Conceptual framework and preliminary

2.1. Conceptual framework

Fig. 1 shows the overall computational framework of this study. First, to collect human safety perceptions of the built environment, we established a survey based on a sample dataset of street view images. We recruited local volunteers in Helsingborg to collect their safety perceptions. They were outfitted with eye-tracking systems when they engaged with the urban landscapes in street view images. Their behavioral data collected was then used to generate human attention heatmaps, which indicate regions within the images that attract significant human focus. Second, following the collection of safety perception data, we applied image segmentation techniques to street view images to detect urban features. These features were then correlated with the human attention heatmaps generated by eye-tracking systems (Yang et al., 2024). To facilitate a more nuanced analysis, we introduced two metrics including Mean Object Ratio in Highlighted Regions (MoRH) and Mean Object Hue (MoH), which allowed us to identify urban features that significantly capture human attention, thereby enabling us to assess their impact on safety perceptions effectively. After that, we employed a well-established workflow that integrates street view images with computer vision techniques to measure safety perceptions. In particular, we leveraged eXplainable Artificial Intelligence (XAI) models and generated XAI-based heatmaps that highlight the importance of urban features while measuring safety perceptions using AI. By comparing these human attention heatmaps and XAI-based heatmaps, we offer valuable human insights regarding the trustworthiness of safety perceptions measured in prior studies to better align with human behaviors. It is worth noting that this study focuses on safety perceptions, and the proposed framework can be applied to other perceptual dimensions widely studied in urban planning such as liveliness and beauty.

2.2. Preliminary: Measuring safety perceptions using street view images and computer vision

Before introducing the methodologies employed in this study, here, we outline a common practice in prior place perception studies that combines street view images and computer vision, which forms the foundation of our approach. In this study, we will leverage two distinct models that have been utilized to measure place perceptions, each based on a different dataset. The first model was developed based on the MIT Place Pulse dataset, which provides a global perspective on place perceptions with over 80,000 participants (Dubey, Naik, Parikh, Raskar, & Hidalgo, 2016), hereafter referred to as the *global dataset* and the *global model*. The second model was constructed from a dataset with over 2000 residents in Stockholm, Sweden (Kang et al., 2023), which aligns with the country in this paper, subsequently referred to as the *Sweden dataset* and the *Sweden model*. The Sweden model delivers insights from a contextual perspective in Sweden, as prior studies have indicated that datasets derived from local people, with contextual knowledge, offer a more accurate reflection of place perceptions than the global dataset (Kang et al., 2023; Yao et al., 2019). Combining both models could help provide a preliminary estimation of the safety perceptions from street view images from global and local perspectives.

The development of both models began with the launch of a survey inviting participants to evaluate their environmental perceptions through a set of selected street view images. In these surveys, participants were presented with two randomly selected images and asked to select one of them to respond to survey questions such as "Which place looks safer?" Notably, the MIT Place Pulse dataset collected six dimensions of human place perceptions including safe, beautiful, depressing, lively, wealthy, and boring, while the Sweden dataset collected safety perceptions of the environment. After collecting participants' preferences, we derived and calculated safety perceptual scores from these evaluations as indicators of the participants' impressions of each street view image. These scores were then leveraged as proxies to represent people's general safety perceptions of the surrounding environment. Following this, we trained a Deep Convolutional Neural Network (DCNN) model to decipher the human perception of a place. Upon completion of its training, the model could be utilized to estimate the perceptual score of any new street view image. The human safety perceptual scores were delineated on a scale ranging from 1 to 9, where a higher score indicates a safer environment. Both models were further leveraged to prepare the safety perception dataset, which will be utilized to measure human safety perceptions using eye-tracking systems. To further compare the attention maps generated from eye-tracking systems and XAI models, we performed those XAI methods based on the Sweden model, as it may better reflect residents' perceptions in Sweden (Kang et al., 2023).

3. Measure safety perceptions with eye-tracking systems

3.1. Dataset

We recruited participants and conducted our experiments in Helsingborg, Sweden. To ensure privacy, no identifiable personal information was collected from participants. A street view imagery dataset that includes 300 images was then constructed to represent the urban streetscapes of Helsingborg. Street view imagery captures detailed characteristics of urban environments, from architectural styles and building facades to road networks and green spaces (Biljecki & Ito, 2021; Kang et al., 2020). Given that street view imagery offers a comprehensive visual representation of urban landscapes, it has been used for collecting human safety perceptions under the hypothesis that human reactions to these street view images can serve as effective proxies for their perceptions of the built environment (Kang et al., 2023; Zhang et al., 2021). To create this dataset, we randomly selected images across the city to ensure the representativeness and spatial coverage. We also selected images that evoke people safe and unsafe perceptions to support consequent data collection experiments that integrates eye-tracking systems. By leveraging eye-tracking systems to collect human reactions to the street view imagery, we aim to uncover a more nuanced understanding of how the urban environment influences residents' perception of safety. More detailed data preprocessing steps for preparing the survey are provided in Appendix A.1

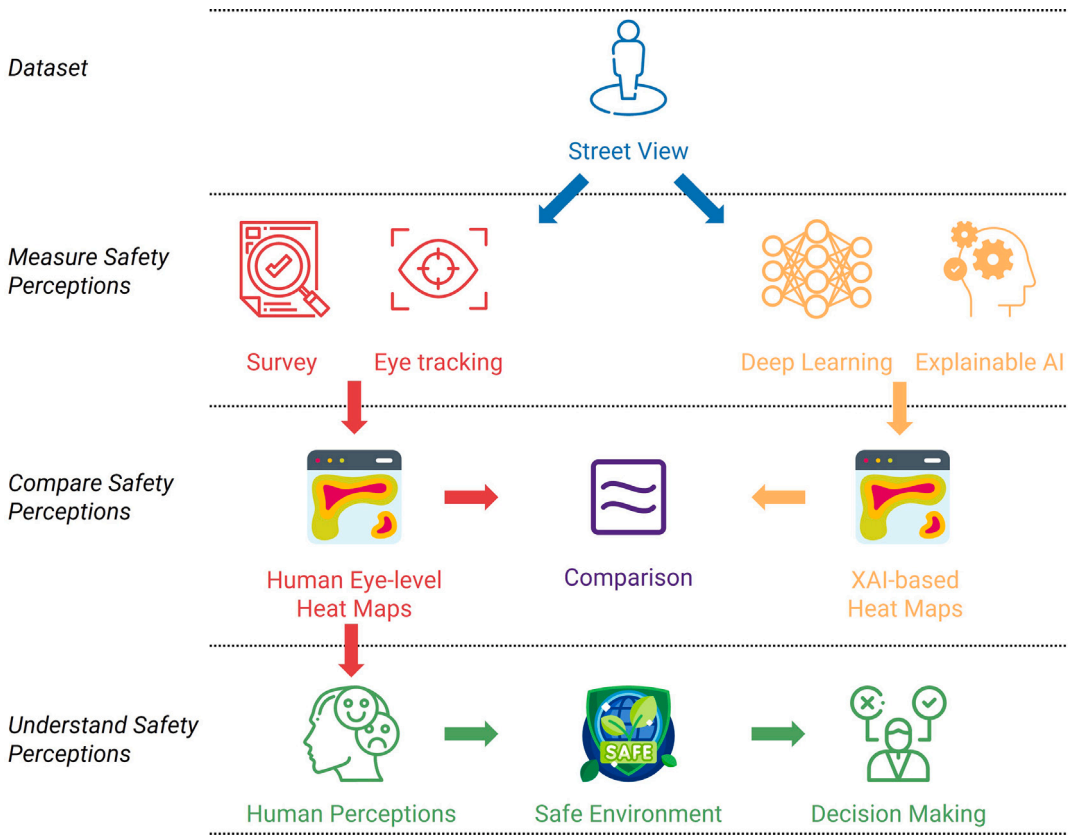


Fig. 1. Conceptual framework of this study. We start from collecting street view images to measure perceived safety. Participants complete a survey and equip eye-tracking systems to produce human attention heatmaps. We also run deep learning-based models to predict safety scores with Explainable AI (XAI) approaches to generate XAI-based heatmaps. By comparing the two results, we aim to deepen our understanding of urban safe environments to inform decision-making for urban planning and design.

3.2. Survey setup

Based on the street view imagery dataset which contains 300 sample images, we launched a local online survey in Helsingborg and integrated eye-tracking systems to collect and understand safety perceptions (Kang et al., 2023). We deployed eye-tracking systems in four places in Helsingborg. We utilized two models of eye-tracking systems, including the TobiiX3-120, offering 0.5 degrees of precision at 120 Hz, and the Tobii-4c, also with 0.5 degrees of precision but at 60 Hz, across a standard distance of 70 cm. Both devices were operated under a researcher's license, ensuring high-quality data collection. The data from the TobiiX3-120 was downsampled to 60 Hz to standardize the output. Using Tobii's proprietary software, blinks were automatically filtered out, and both fixations and saccades were accurately identified through the integrated iVT-Tobii filter.

We adopted eye-tracking systems as a core place perception assessment tool in this study to complement traditional survey methods. Eye-tracking systems could capture participants' visual attention when looking at images, offering insights into how individuals visually engage with different urban elements. Unlike traditional methods such as questionnaires or semantic differential scales, which rely on subjective recall and post-hoc self-reporting, eye-tracking systems capture human gaze behaviors, enabling the measurement of more nuanced safety perceptions. We invited passersby to share their safety perceptions in response to street view images. Overall, responses from 127 participants were collected from this survey. On average, each street view image was compared 5.8 times with others. We analyzed the demographic information of our participants. In total, we received valid responses from 127 participants. The gender ratio (male vs. female) is 1.25:1, with White individuals comprising over 93% of the population.

Fig. 2 shows the user interface of the created survey. Each participant was first asked to provide their demographic details, including age and gender, for further analysis of potential population biases. Then, participants were exposed to ten pairs of two random street view images and asked to select, for each pair, "Which place looks safer?". Each participant was equipped with an eye-tracking system so that their gazes when viewing the street view images were recorded. By tracking participants' eye movements, we could identify which objects within the images attracted their attention, triggered their decisions, and therefore shaped their safety perceptions. We followed the established workflows of surveys from prior studies to ensure consistency in our comparisons between human eye-level safety perceptions using eye-tracking systems, and AI-generated safety perceptions with XAI methods (Dubey et al., 2016; Kang et al., 2023; Zhang, Zhou et al., 2018).

3.3. Human attention heatmaps

The eye-tracking data, collected from participants in evaluating safe environments, was then translated as visual heatmaps. These visual heatmaps obtained from eye-tracking systems are further termed *human attention* heatmaps. Visual heatmaps have been utilized as a common method for visualizing human attention through eye-tracking systems, as they effectively illustrate the focal areas observed by participants (Kiefer et al., 2017; Raschke, Blascheck, & Burch, 2014). Eye-tracking systems capture the duration and locations of gaze fixations across street view images. Based on this, we aggregated all gaze points of each pixel and performed a cumulative distribution function to represent the accumulated focus intensity for that pixel to create the heatmap (Raschke et al., 2014). It should be noted that this study only

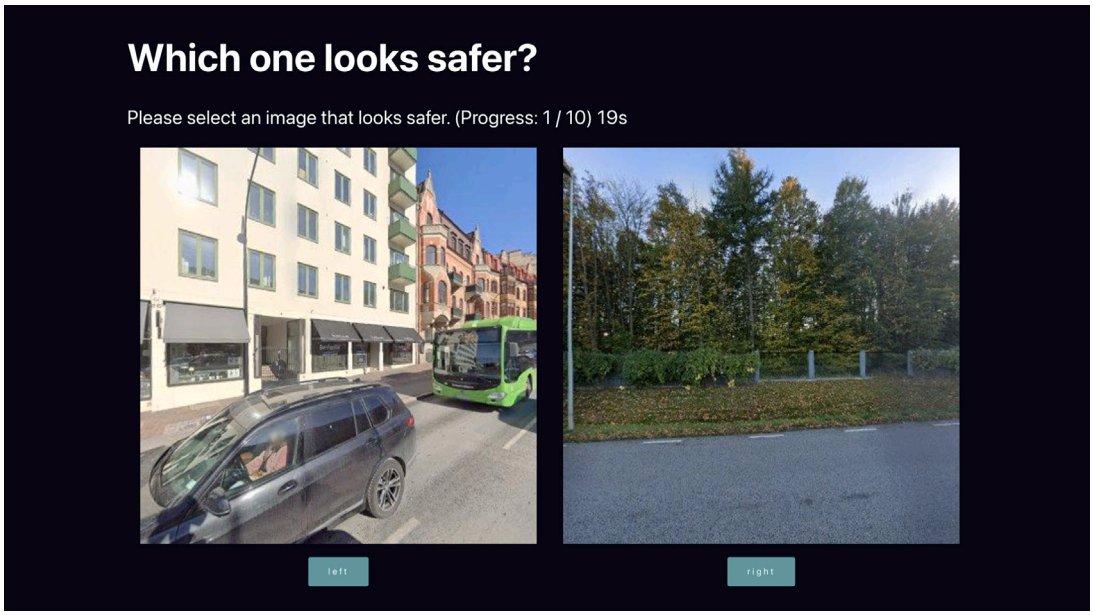


Fig. 2. A user interface screenshot of the survey.

leveraged aggregated spatial heatmaps of images to capture patterns of visual attention across images. Temporal aspects of gaze behavior were not modeled in this analysis. As illustrated in Fig. 7, the heatmaps use a color gradient to indicate the density of gaze attention: warmer colors such as reds and yellows highlight areas of interest (AOIs) that attracted more observation, whereas cooler colors represent AOIs that received less visual attention. Specifically, our analysis utilizes the hue scale as a metric for this spectrum of attention. While raw hue values range from 0 to 360 degrees, we applied a linear rescaling transformation to map this spectrum onto a 0–150 range purely for visualization purposes. Regions with lower hue values (red/yellow, warmer colors) reflect high visual attention, and regions with higher hue values (e.g., blue, cooler colors) reflect lower human attention. This transformation did not affect any computational analyses and data distributions, and facilitates the intuitive interpretation of gaze intensity and attention distribution across street view images. It also enables us to further identify features in urban streetscapes that potentially impact human safety perceptions. Moreover, it allows for a comparative analysis between heatmaps generated from eye-tracking systems and those generated using XAI models.

After translating images into heatmaps to represent human visual attention, we further categorize street view images into two groups based on participant selection frequencies in the pairwise safety comparison task. Specifically, we classified images into a "more frequently selected as safer" group (safe) and a "less frequently selected as safer" group (unsafe). Images classified as *safe* were those consistently chosen as safer at least three times, whereas *unsafe* images were those not chosen as safer images at least three times. These labels are derived from aggregated human choices, rather than from model scores, and reflect behaviorally evoked safety perceptions during comparison. By doing so, we could further link these behavior-grounded safety perceptions with urban features that draw human attention, thereby shedding light on modeling human–environmental interactions.

3.4. Image segmentation for MoR

We further extracted objects from street view images using image segmentation methods, aiming to analyze what elements and objects may trigger human safety perceptions. To do so, we leveraged a cutting-edge deep learning method based on Vision Transformers (ViT), utilizing the Dense Prediction Transformer (DPT) model. The DPT model is trained on the ADE20K dataset and could identify 150 common objects in urban places (Ranftl, Bochkovskiy, & Koltun, 2021; Zhou et al., 2017). This image segmentation pipeline has been widely applied in recent place perception studies (Ceccato et al., 2025; Ma, Zhang, Cui, Kwan, & Cai, 2025). As illustrated in Fig. 7, urban features were identified with different colors in the example street view image. We can quantify their proportions in street view images using the following equation:

$$R_o = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w}\mathbb{1}(p(i,j)=o)}{h\times w} \quad (1)$$

where R_o indicates the proportion of the object o within an image, and $p(i,j)$ refers to the pixel value at the row i and column j of the image. It is computed by dividing the number of pixels corresponding to object o by the total number of pixels in the street view images (h and w are the height and width of images). By computing the mean values of $\overline{R_o}$ (Mean Object Ratio, MoR) across all images for each object, we could measure the frequency of urban features in urban landscapes captured in the street view imagery.

3.5. Linking attention heatmaps for important feature identification using MoRH and MoH

Performing image segmentation to identify urban elements from street view images and build associations between urban features and perceptions has been a common practice in existing literature (Larkin et al., 2021; Ramírez et al., 2021; Zhang, Zhou et al., 2018). However, as illustrated in Fig. 7, when viewing street view images, participants may have different focus and attention on certain objects. Therefore, it is necessary to take people's focus into account. To quantify what objects are important in shaping and affecting human safety perceptions, we propose two metrics: Mean Object Ratio in Highlighted Regions (MoRH, $\overline{RH_{(t,o)}}$) and Mean Object Hue (MoH, $\overline{Hue_o}$).

The first index, *Mean Object Ratio in Highlighted Regions* (MoRH, $\overline{RH_{(t,o)}}$), is calculated to quantify the frequency of objects that appear in the most visually attention AOIs of an image, as defined by a specified

threshold t. Different from the Mean Object Ratio (MoR), which considers the average presence of objects throughout the entire image, the Mean Object Ratio in Highlighted Regions (MoRH) specifically focuses on those areas receiving the highest levels of human attention, thereby quantifying the appearance of objects in these AOIs. Considering the variance in the size of these highlighted AOIs, the MoRH metric is normalized by the number of pixels of these AOIs within the image. The MoRH is expressed as follows:

$$RH_{(t,o)} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w}\mathbb{1}(p(i,j)=o \wedge Hue(p(i,j)) \le t)}{\sum_{i=1}^{h}\sum_{j=1}^{w}\mathbb{1}(Hue(p(i,j)) \le t)} \quad (2)$$

where $RH_{(t,o)}$ indicates the proportion of the object o within the highlighted areas, characterized by hue values falling below the threshold t.

The second index, *Mean Object Hue (MoH)*, calculates the average hue value of a certain object in street view images. By using the hue value to depict the intensity of human visual focus, the MoH value could indicate the level of attention directed towards a specific object. Compared with MoR, MoH takes hue values as weights to indicate the importance of different urban features. In our study, the hue values in heatmaps range from 0 to 150. It is worth noting that we adjusted the MoH by computing $150 - MoH$ to get its reverse meaning for easier interpretation. This ensures that a higher MoH, like 150, corresponds to a strong level of attention, and 0 represents a low level of attention to a particular object. By proposing the two metrics mentioned above, including MoRH and MoH values, we could quantify the importance of objects in street view images when participants perceive the images, rather than treating them equally.

4. Understand safety perceptions

To discover what street elements trigger human safety perceptions, we first analyzed street features by comparing the proportions of objects in street view images with Mean Object Ratio (MoR) and solely in highlighted areas using Mean Object Ratio in Highlighted Regions (MoRH). After that, we identified important street elements based on Mean Object Hue (MoH).

4.1. Object detection without eye-tracking systems

Adopting methodologies utilized in prior studies, we employed image segmentation to quantify the occurrence of various objects in street view images. We aim to reproduce prior practices to assess perceived urban safety without integrating eye-tracking systems. Given that we have divided street view images into two groups based on participant selection frequency: safe and unsafe, we identified the top 10 objects most commonly present in each group, as depicted in Fig. 3. Five key categories of street elements were identified, each contributing significantly to the composition of urban streetscapes: *Architectural features* such as buildings, ceilings, and bridges are frequently observed in street view images for creating urban streetscapes. *Pathways* including roads, paths, and sidewalks, are commonly seen in street view images, with sidewalks having more appearances in "safe images", indicating a potential positive correlation between pedestrian pathways and sense of safety. *Urban greenery*, such as trees and grass, appearing frequently in street view images, highlights the significance of green spaces in urban settings. Variables such as sky and ground that influence the *openness* and the overall atmosphere of urban scenes are also key elements in street view images. In addition, the presence of *vehicles* such as trucks and buses, often observed in "unsafe images", implies a possible negative association between vehicles and safe environments. Overall, these observations are consistent with several findings from prior studies. For example, elements such as buildings, trees, and sky typically occupy substantial proportions of pixel space in street view imagery (Li et al., 2022; Ramírez et al., 2021). Urban greenery and open spaces are typically associated with positive environmental perceptions (Ramírez et al., 2021; Zhang, Zhou et al., 2018). While the presence of vehicles might have a negative impact on safe environments. It is worth noting that the occurrence of street elements in images does not necessarily directly reflect their impacts on safety perceptions. It emphasizes the complex human–environment relationships, highlighting the significance of delving into such interactions to better understand how various tangible street elements influence human subjective safety perceptions.

4.2. Object detection with eye-tracking systems

By incorporating eye-tracking systems into our study, we can detect specific urban elements that attract human attention when observing street view images. Fig. 4 presents several representative images from the safe and unsafe groups together with aggregated human attention heatmaps from participants. This overlay of attention maps supports a more nuanced understanding of the human behaviors and attention that shape urban safety perceptions. To associate human gazes with urban elements, we performed image segmentation and computed the Mean Object Ratio in Highlighted Regions (MoRH) and Mean Object Hue (MoH) to identify important urban elements.

In our analysis, it is critical to acknowledge that the previous results in Section 4.1 have given equal weights to all objects within street view images, potentially overshadowing the relative importance of some objects in attracting human attention. Fig. 5 illustrates the top 10 objects ranked by their MoRH in the two image groups (safe vs. unsafe). The frequency of streetscape objects is calculated by performing image segmentation on the highlighted regions. We evaluated two thresholds t, 15 and 30, to compute the MoRH, as we do not have a precise threshold. We then analyzed the most common street elements in those highlighted regions.

Comparing the results in Section 4.1, we found notable discrepancies between the outcomes generated by eye-tracking systems and those obtained without this technology. For images in the safe group, signboards, cars, trucks, and plants were identified as more significant based on eye-tracking data which aligns with prior studies (Dong et al., 2023). In terms of key street elements associated with images in the unsafe group, we found that trucks, vans, cars, houses, doors, and plants are more significant compared to results obtained without eye-tracking systems. Conversely, ceilings were not among the top 10 objects in areas with the most human attention when $MoRH(t = 15)$ in safe images, yet they were ranked among the top three objects when $MoRH(t = 30)$. This implies that they take up a considerable portion of the image but may not be the most important features in influencing perceived safe environments. Additionally, the eye-tracking data showed that the sky and pathways, such as paths and sidewalks, were perceived as less significant in both groups of images. Such interesting findings demonstrate that sky may not be a focal variable in shaping human safety perceptions, different from prior work (Ogawa et al., 2024). This comparison provides a detailed analysis of how humans visually perceive and prioritize urban elements using eye-tracking technology. It emphasizes the importance of exploring the nature of human visual perceptions rather than solely focusing on the occurrence of objects in images.

Next, we computed the Mean Object Hue (MoH) to analyze the attention intensity at the element level, considering the varying sizes of urban objects and elements. Fig. 6 presents the top 10 objects ranked by MoH in images from both safe and unsafe groups. We observed significant discrepancies in results when human attention was considered in identifying important street elements, which highlights the complex relationship between physical urban elements and subjective human perceptions. *Transportation vehicles* such as buses, vans, and cars attracted more human focus, particularly in unsafe images, suggesting an association with unsafe environments. *Urban infrastructure* elements draw more human attention to both safe and unsafe images. Stairways, floors, bases, and railings, exhibit a positive

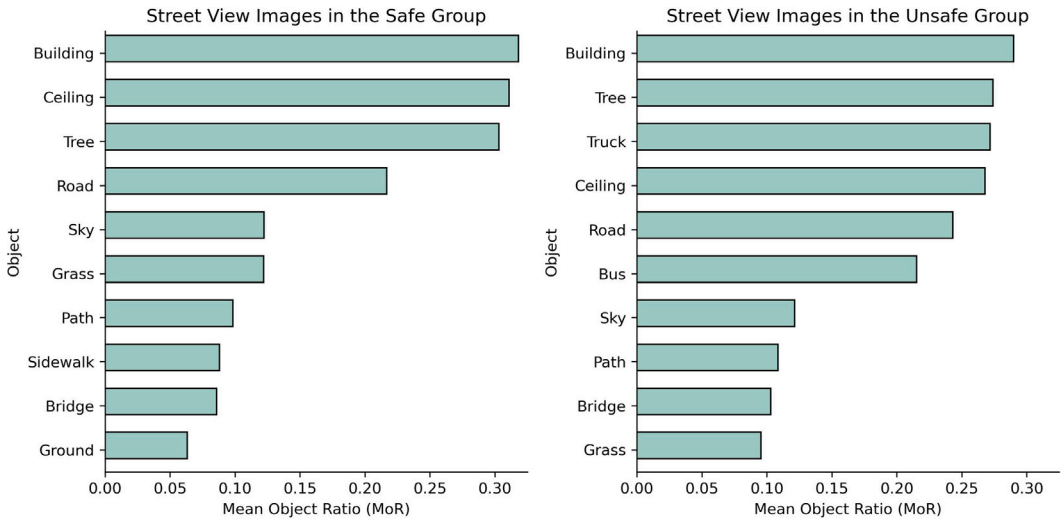


Fig. 3. Top 10 objects based on Mean Object Ratio (MoR) in the two groups of street view images.

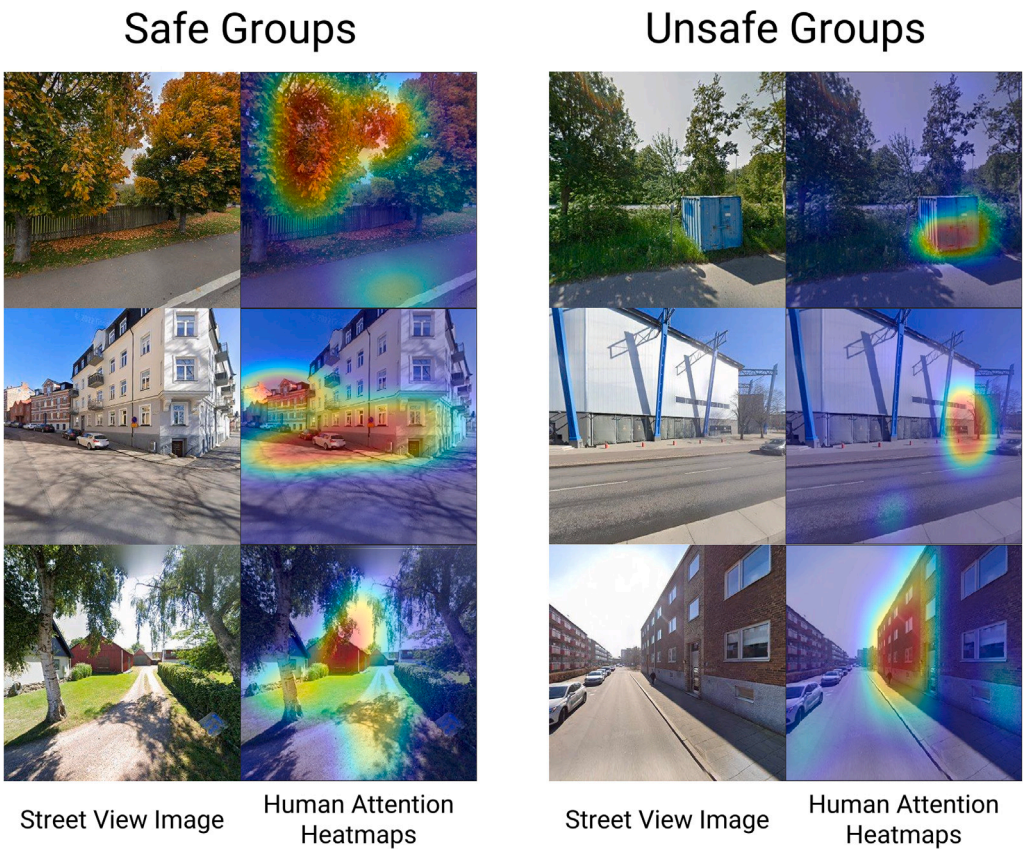


Fig. 4. Sample street view images in safe and unsafe groups with aggregated human attention heatmaps.

correlation with safety perceptions, highlighting their role in creating secure urban places. while elements like railings, signboards, fences, and hovels might be associated with negative safety perceptions, indicating potential barriers in urban environments. *Public space* features such as booths, chairs, flags, plants, and boxes also have high MoHs, and might be positively associated with safety perceptions, reflecting their roles in enhancing the quality and usability of public spaces. Land and sea, as *natural elements*, have shown negative associations with safety perceptions. This aligns with prior work suggesting that open natural elements can evoke a sense of vulnerability due to their lack of enclosure, restricted visibility, and human presence (Nasar & Jones, 1997; Stamps III, 2005). Environments with limited opportunities for surveillance, reduced escape routes, and isolated environments might be associated with increased perceptions of unsafety (Appleton,

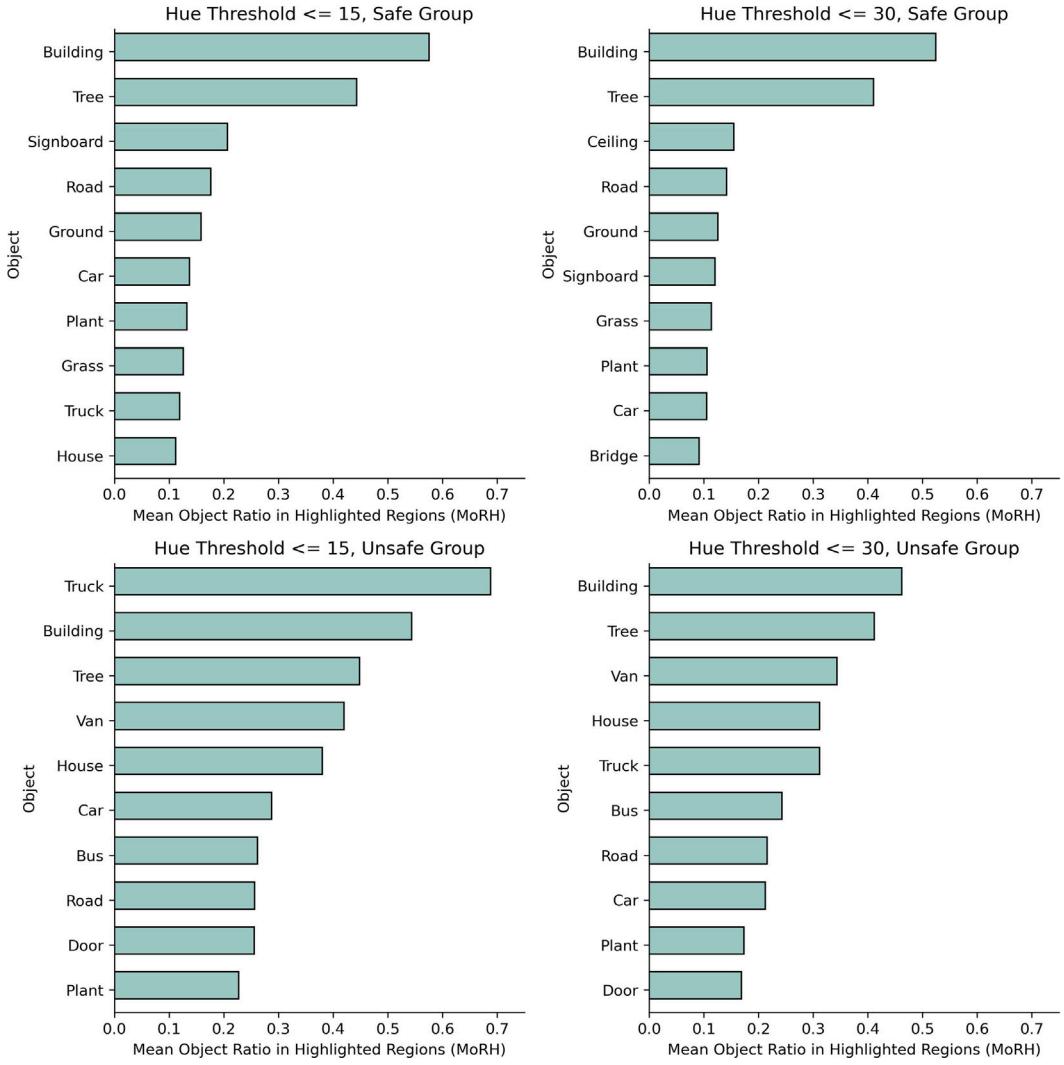


Fig. 5. Top 10 objects based on Mean Object Ratio in Highlighted Regions (MoRH) in the two groups of street view images.

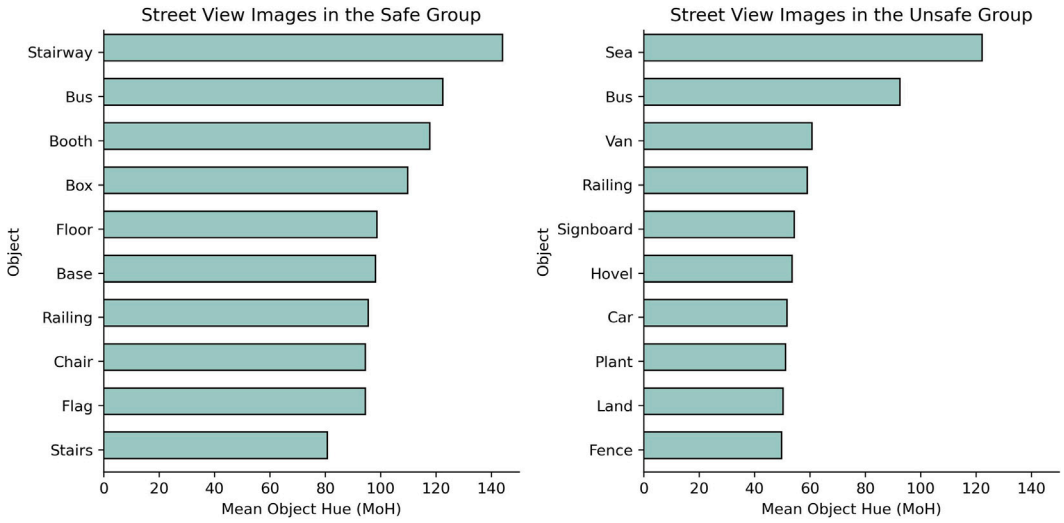


Fig. 6. Top 10 objects based on Mean Object Hue (MoH) in the two groups of street view images.

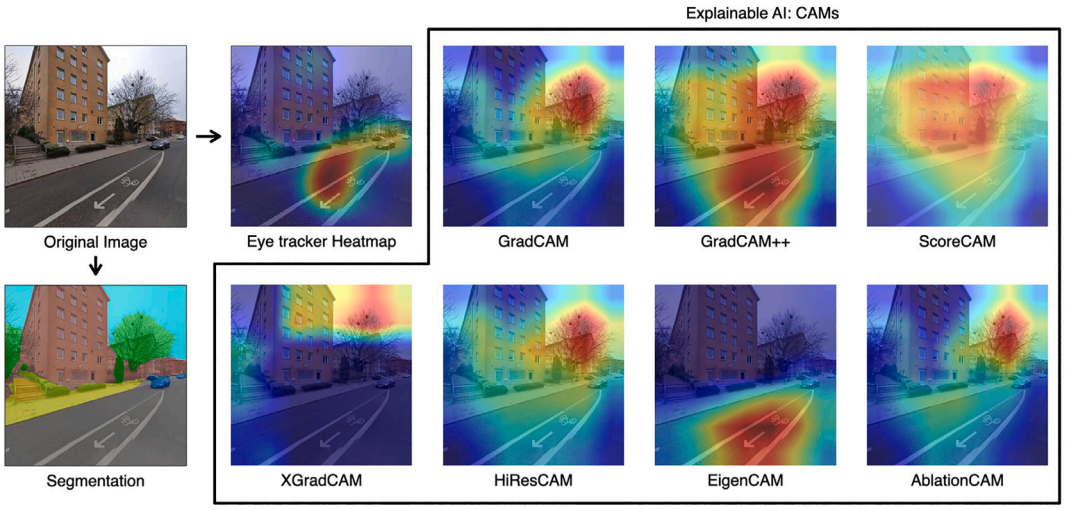


Fig. 7. Examples of the original street view images, image segmentation results, human visual attention, and a series of heatmaps.

1975; Fisher & Nasar, 1992). The results mentioned above not only demonstrated the crucial role of delving into human perceptions with eye-tracking systems but also offer new insights into illuminating the nature of the safe-environment nexus.

5. Safety perception comparison

This section aims to assess whether place perceptions measured using XAI methods could accurately reflect human perceptions. Thus, we compare heatmaps generated by XAI techniques with those produced from human eye-tracking data.

5.1. Explainable Artificial Intelligence (XAI)

In prior studies, researchers have leveraged XAI methods such as Gradient-weighted Class Activation Mapping (GradCAM) to examine the associations between safety perceptions and built environments (Li et al., 2022; Moreno-Vera, 2021). These methods help understand which elements in urban environments play important roles in shaping human subjective place perceptions. In our study, we also followed this strategy and utilized Class Activation Mapping (CAM) to identify objects in street view images that significantly influenced the classification outcomes from the deep learning model. In particular, we utilized CAM to create heatmaps from the Sweden model, as it may better reflect Swedish safety perceptions.

Deep Convolutional Neural Networks (DCNNs) contain multiple convolutional layers. Each of these layers produces feature maps that capture different patterns of the input image, such as edges, textures, or shapes. These feature maps are essential as they highlight the specific characteristics that the model identifies, and are thereby utilized by CAM approaches to provide explanatory insights. In particular, we leveraged seven CAM techniques, including GradCAM, ScoreCAM, GradCAMPlus, AblationCAM, XGradCAM, EigenCAM, and HiResCAM. Each of these offers nuanced visualizations of influential regions within the input image, represented as heatmaps. These approaches vary in their focus and characteristics. GradCAM emphasizes the gradients flowing into the last convolutional layer; ScoreCAM recalculates class scores, considering individual feature maps; GradCAMPlusPlus extends GradCAM by considering the positive and negative impacts of the feature maps; AblationCAM ablates each feature map and measures the output change; XGradCAM introduces a modified formulation to produce more visually sharp regions; EigenCAM leverages principal component analysis; and HiResCAM focuses on providing high-resolution maps. The derived heatmaps showcase areas that have a substantial impact on the classification of safety perceptions. The presence of warm hues (ranging from vibrant yellows to intense reds) within heatmaps indicates regions where the observed urban elements are more important in shaping subjective safety perceptions. An example of these heatmap outputs can be seen in Fig. 7, demonstrating the practical application and visual impact of our selected CAM techniques. Since we have already obtained the human attention heatmaps, to enable comparison, we standardized the hue values of these XAI-based heatmaps between 0 and 150 as well. After that, both groups of heatmaps were input into the image segmentation model to identify important objects and to further compare their differences.

5.2. Image similarity comparison

After we obtained the two groups of heatmaps, including human attention heatmaps and XAI-based heatmaps, we quantified their similarities to help understand the reliability and trustworthiness of XAI models and identified which one aligns most closely with human vision. To accomplish this, we comprehensively measured the image similarity at two levels: *scene* level and *element* level (Kang et al., 2020).

At the scene level, we aim to quantify the overall similarity between any two images. Two metrics, the L2 loss, and the Learned Perceptual Image Patch Similarity (LPIPS) score (Jang et al., 2024; Zhang, Isola, Efros, Shechtman and Wang, 2018), were computed between the two groups of heatmaps. Both metrics range from 0 to 1 and assess the overall similarity between images. A lower L2 loss value indicates a greater similarity between human and XAI heatmaps; while lower LPIPS values indicate greater similarity between human and XAI heatmaps. By using these two measures, we measured the image similarity between the human attention and XAI-based heatmap images to identify the XAI model that best matches human visual attention. More technical details about these two metrics are provided in Appendix A.2.

At the element level, we focused on verifying whether the objects that received people's attention in both sets of heatmaps were consistent. To accomplish this, we first calculated the MoH value for each object across all heatmaps. Each heatmap is depicted as a 150-dimensional vector. We then compare the similarity between vectors of human attention heatmaps and their corresponding seven XAI-based heatmaps by calculating the cosine similarity. The hypothesis is that objects that appeared in both heatmaps should be similar in two groups of images. By doing so, we could determine which specific XAI model most aligns with human perceptions, offering valuable insights into the effectiveness and reliability of different XAI models when understanding human safety perceptions.

Table 1
Image similarities between human attention heatmaps and seven XAI-based heatmaps based on three measures. Top 2 most similar XAI methods are in bold.

Model name	Loss ↓	LPIPS score ↓	Cosine similarity ↑
AblationCAM	0.4232	0.5855	0.0132
EigenCAM	**0.3512**	**0.5478**	**0.0143**
GradCAM	0.4357	0.5896	0.0130
GradCAMPlusPlus	0.4998	0.5891	0.0100
HiResCAM	0.4386	**0.5688**	0.0127
ScoreCAM	0.4631	0.5806	0.0100
XGradCAM	**0.3285**	0.5739	**0.0161**

5.3. Comparison results

To answer the second research question, we evaluated the trustworthiness of using street view images and deep learning to measure safety perceptions at both scene and element levels. At the scene level, we compared the image similarities between heatmaps generated by XAI models and those derived from human assessments utilizing eye-tracking systems. Both the L2 loss function and LPIPS scores were computed. The comparative analysis of these XAI models using the two metrics in Table 1 demonstrates similar trends. XGradCAM and EigenCAM have the lowest L2 loss values of 0.3285 and 0.3512, respectively. They also ranked in the top 3 models with the lowest LPIPS scores of 0.5739 and 0.5478 (the lowest), meaning they have high similarities with heatmaps based on eye-tracking systems. This indicates that the outputs from these two XAI approaches trained based on the Stockholm model closely match the results from using human eye-tracking systems. If we treat the human eye-tracking results as the benchmark for measuring human safety perceptions, these findings suggest that XGradCAM and EigenCAM may offer greater reliability and might be more trustworthy when using XAI models for future applications. It also emphasizes the significance of involving humans in the process of developing and evaluating explainable artificial intelligence to derive interpretable insights.

At the element level, we first convert the proportions of objects into a 150-dimensional vector and then measure the appearance of the two vectors. Results of the cosine similarly show similar findings, in comparison with the results at the scene level. Both XGradCAM (0.0161) and EigenCAM (0.0143) have the highest cosine similarities, indicating the objects that appeared in human attention heatmaps are similar to those objects in these two XAI-based heatmaps. It implies that these two XAI models have similar patterns with human eye-level perceptions and might be more reliable in explainable objects that trigger safety perceptions.

To gain further insight into contexts where XAI-based heatmaps diverge from human attention, we examined a few representative street view images with relatively low cosine similarity and high LPIPS values. These images have relatively weak alignment between XAI-generated and eye-tracking heatmaps, as illustrated in Fig. 8. We selected these images from the bottom quartile of similarity scores, focusing on EigenCAM, as it shows the highest alignment across our experiments.

We observed two general patterns. First, while EigenCAM successfully identified objects that are associated with safety perceptions, it tended to produce broader and more diffuse heatmaps, highlighting larger regions. In comparison, human attention maps show more selective and focused patterns, highlighting a smaller number of pixels (Fig. 8(a), (b), (e), (f)). For instance, in Fig. 8(a), (b), and (e), EigenCAM highlighted a wide area including multiple vehicles, whereas participants tended to focus only on one or two vehicles directly related to perceived safety. This discrepancy likely reflects the selective nature of human vision to concentrate on a small number of visual elements, whereas CAM techniques highlight more generalized regions. These examples illustrate that low similarity may arise even when both heatmaps highlight the same objects. Second, it is possible that the semantic focus of EigenCAM-based heatmaps is notably different from that of human observers. For example, in Fig. 8(g) and (h), participants focused on facade textures, while CAM-based heatmaps highlighted pathways. Our findings highlight notable limitations in current XAI techniques. When leveraging AI for urban planning practices, models that do not align with human perception could lead to misleading conclusions. Therefore, we advocate for the need to develop human-aligned XAI methods that better reflect human perceptions. Additionally, it is necessary to be cautious when interpreting XAI outputs in supporting urban decision makings.

6. Discussions

We discuss several takeaways from this paper below.

6.1. Associations between safety perceptions and built environment

This study deepens our understanding of how the built environment influences perceived safety by incorporating human gaze behavior through eye-tracking systems. In contrast to traditional computer vision approaches that treat all objects equally, our results reveal that human attention is selectively drawn to micro-level features. This aligns with the psychological principle of selective attention (Lavie, Hirst, De Fockert, & Viding, 2004). We observed that urban features such as stairways, floors, and public space amenities, such as booths, flags, boxes, might be associated with safety perceptions, as they may enhance visibility and active use of places, While vehicles and objects indicating potential barriers such as fences and hovels might be associated with negative safety perceptions. Our findings complement and extend prior studies. For example, Dong et al. (2023) highlighted natural views and amenity-related features in influencing positive urban behaviors, which aligns with our findings to enhance safety perceptions. While we have also noticed some discrepancies as sky might attract less attention, which is different from prior work (Ogawa et al., 2024). Our study contributes to the literature to move from "what is present" to "what people look at", offering actionable design guidance for building safer neighborhoods.

A key finding is the mismatch between the visual size of objects in street view images and the amount of attention they receive, as displayed in human attention heatmaps. Though occupying more pixels in images, elements like the sky or ceilings attract significantly less attention. Conversely, certain features receive disproportionate visual focus, especially those related to navigation, rest, or human activity. These findings support classic theories in environmental design and criminology. For instance, the prominence of micro-scale elements in "safe" images echoes Newman's Defensible Space (1973), which emphasizes territorial cues such as thresholds and semi-private spaces that enhance perceived control and ownership. Similarly, our results align with principles from Crime Prevention Through Environmental Design (CPTED) (Jeffery, 1971), which advocate for natural surveillance, territorial reinforcement, and spatial legibility as key strategies to reduce crime and enhance safety perception. The frequent association of vans and large vehicles with perceived "unsafe" environments reflects the logic of the Broken Windows Theory (Wilson James & Kelling George, 1982), where visual signals of disorder and transience may erode one's sense of security. Additionally, stairways and railings contribute to perceived safety by offering vantage points that are associated with the core Prospect-Refuge Theory (Appleton, 1975). Symbolic or identity-linked elements, like signage and flags, illustrate how urban legibility and recognition influence perception. These insights offer useful references for urban design and planning practices. Rather than focusing solely on large-scale openness or aesthetic uniformity, designers may consider incorporating perceptually salient features that enhance spatial legibility and support informal territorial cues. Especially in the case of pre-designed or existing environments, these findings could inform visual simulation-based evaluations before

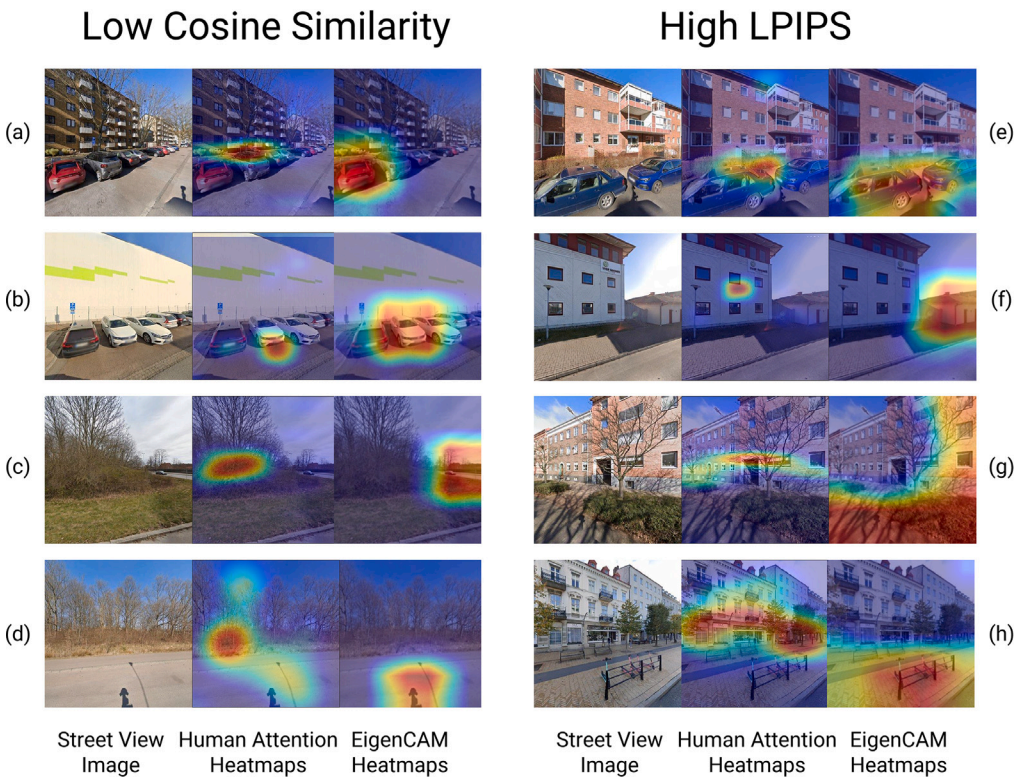


Fig. 8. Representative examples of low alignment between XAI-based and human attention heatmaps. Images (a–d) show images with low cosine similarity values, and images (e–h) show images with high LPIPS values.

real-world implementation, allowing for perception-informed feedback loops.

Some limitations warrant further discussion. While our analysis reveals how certain physical features influence gaze behavior and perceived safety, the concept of "safety" may vary across contexts and individuals and is also intertwined with social and situational environments. Despite that fear and safety perceptions are used interchangeably in this manuscript, they have different meanings. Fear refers to an affective or emotional state, and safety perceptions reflect individuals cognitive judgments of environments. In our study, the question "Which place looks safer?" may elicit more cognitive evaluations, rather than internal psychological distress. The interpretations of safety may vary across individuals, as someone may focus on crime-related concerns, while others may consider traffic safety or environmental hazards. Furthermore, social contexts play an important role in shaping safety perceptions. Factors such as the presence of other people, the topological relationships among urban elements, the perceived possibility of being seen by others, as emphasized by "eyes on the street" proposed by Jacobs (1961), are crucial in shaping the sense of safety. Our current methodology does not fully capture these individual differences and social dimensions, and focuses primarily on visual stimuli. Despite these limitations, our study contributes to the existing body of research on safety perceptions and place perception by introducing more human experiences with psychological methods. To provide a more comprehensive understanding of safety perceptions in urban contexts, more behavioral data that integrates affective, cognitive, and social aspects of place perceptions is necessary in future studies.

6.2. Advanced technology for environmental perception

This paper serves as an example of integrating emerging datasets and technology with environmental psychology to enrich our understanding of the built environment. Using an interdisciplinary approach that combines emerging datasets and technologies, we could measure subjective human experiences at place in response to physical built environments. On the one hand, we have demonstrated the opportunities of utilizing geospatial data science to obtain deeper insights into environmental perception studies and the complex interactions between the built environment and human perceptions. Using street view images and deep learning approaches, we could characterize and model the built environment effectively and efficiently. On the other hand, the use of several advanced psychological methods, such as eye-tracking systems, opens new avenues for decoding the nature of how individuals interact with their surrounding environments. We advocates for more human-centered insights. By combining the two approaches together, this paper contributes to a more holistic view of environmental perception, offering a comprehensive framework that potentially could guide both research and practical applications in urban planning and public policy.

Beyond safety perceptions, the integration of advanced technologies also brings more opportunities for the future environmental perception studies. Currently, advanced AI algorithms, like those supported by Large Language Models (LLMs), can now communicate with people in a natural way and even enrich our understanding of human subjectivity and feelings. For instance, Jang et al. (2023) have leveraged LLMs to explore place identity across global cities. Additionally, emerging psychological techniques such as fMRI (functional magnetic resonance imaging) and brain–computer interfaces have provided fresh insights into human cognition and perceptions (Yang et al., 2025; Zhao, Feng, Sun, Chang, & Shaw, 2024). This study demonstrates how combining the two types of advanced technologies could facilitate our understanding of environmental perceptions. In addition, the emergence of virtual reality (VR) and augmented reality (AR) technologies, such as the Apple Vision Pro and Meta Quest may offer new opportunities for modeling human–environment interactions (Yang et al., 2025). Leveraging these emerging technologies may further advance our knowledge of the digital world, beyond our physical environment. We advocate for further

interdisciplinary studies to be performed in the future to analyze the complex human–environment interactions.

Equally important, it is also necessary to acknowledge that using eye-tracking systems has challenges as well. Due to the operating complexity, time, and cost when deploying eye-tracking systems, the sample size and geographic coverage are limited. Therefore, a mixed-method design could significantly inform our understanding of the human–environment interactions: using geo-big data and AI models could estimate large-scale human place perceptions, and targeted eye-tracking subsamples could help uncover deeper mechanisms of the human–environment relationships. Given that different methods have their pros and cons, it is necessary to consider their nuances when implementing them in real-world practices.

6.3. Reliability and ethical implications of AI approaches

More importantly, we evaluated the trustworthiness of using street view images and computer vision methods for measuring human safety perceptions. Deep learning and AI approaches have long been criticized due to their "black-box" nature. Researchers have developed a series of XAI models to reveal the underlying mechanisms of AI, but the reliability of these approaches has not yet been thoroughly validated. Alarmingly, prior studies have already treated XAI outcomes as ground truth outcomes. However, with numerous XAI approaches based on a variety of hypotheses, which XAI model is the most trustworthy and reliable? This paper involves human-in-the-loop comparisons between findings from eye-tracking systems and XAI models. Through a critical analysis, we delve into the ethical considerations associated with deploying emerging technologies in urban studies. We not only measure the similarities between human perspectives and AI perspectives but also emphasize the significance of maintaining a human-centric approach when leveraging emerging technologies. It is worth noting that, due to the complex process of AI models, different XAI models may have varied performances across applications and contexts. Our study could represent a start towards encouraging more human-centered approaches to be leveraged for critically evaluating and validating the trustworthiness of the method before integrating them into urban planning practices. This helps build trust and responsible AI applications, leading to the generation of valuable insights to enhance our understanding of safe environments and human–environment relationships more broadly.

Furthermore, the advent of deep learning could be traced back to the invention of perceptron, designed to mimic human cognitive behaviors (Raiko, Valpola, & LeCun, 2012). Our findings indicate that there is still gaps in the capabilities of existing XAI methods to accurately infer human behaviors. While these XAI methods can offer valuable insights, the applications of these findings need to be approached with caution before full integration into practices. We advocate for integrating more human-centered insights, especially from the fields of neuroscience and psychology, to facilitate the development and assessment of AI methods (Ye et al., 2025; Zhao et al., 2024). Incorporating these perspectives may enhance the design of AI systems to more closely reflect actual human behaviors, thereby increasing their reliability and trustworthiness.

6.4. Limitations and future directions

There is still some room for further enhancement of the proposed experiment. One issue refers to the sample size and geographic representativeness of the dataset. It is resource-intensive to recruit participants to wear eye-tracking devices to collect experimental data, which limits the sample data size. As a result, the current analysis does not support comparisons across demographic subgroups, such as gender or socioeconomic background (Zhou et al., 2025). Also, we have only deployed studies in one city and have merely focused on safety perceptions. Future experiments might be performed across different areas to provide more generalized results and outcomes, and adapted to other perceptual dimensions commonly studied in urban studies, such as liveliness and beautiful. Also, future studies may collect pre-annotated image datasets that indicates safety perceptions from other social media platforms to supplement street view images to enhance the generalizability of this study.

We also recognize the need to integrate more psychological knowledge into this study and more geography and urban studies. There have been limited studies at the intersections between place perception studies and psychological methods. We acknowledge that eye-tracking experiments may introduce some potential biases in measuring the "sense of safety" in this study. For example, people may tend to focus on large foreground elements or independent elements (Zhang & Liu, 2017). Due to the sample size, it is challenging to analyze detailed gaze trajectories to understand topological relationships among street elements. Integrating temporal perspectives of eye-tracking systems may offer additional insights to describe the nuanced cognitive mechanisms. Our study did not delve into these psychological dimensions. Thus, on the one hand, we call for more interdisciplinary collaborations to strengthen the methodological robustness of using psychological methods in future studies and to facilitate a more comprehensive integration of diverse disciplinary perspectives. Additionally, incorporating more human-centered approaches into geography and urban studies could enrich our nuanced understanding of the complex human–environment interactions to better modeling human subjective experiences at places.

Another important limitation of this study is the lack of consideration for participants' sociodemographic backgrounds. The perception of "safety" is highly subjective and may vary significantly across different population groups. For example, children and older adults may prioritize ease of mobility or visibility of assistance, while people with disabilities might focus on barrier-free access and navigability. Individuals of different genders and from diverse ethnic or cultural backgrounds may interpret environmental cues differently due to prior experiences or systemic inequalities (Zhou et al., 2025). Moreover, the degree of familiarity with the environment, such as whether a person is a local resident or a visitor, can also influence safety perception. Residents may feel more at ease in spaces they know well, even if those spaces appear disordered to outsiders. In our study, over 93% of participants identified as White, such demographic homogeneity may limit our ability to generalize findings to more diverse populations. Given that our study did not systematically account for these variables, future research may aim to incorporate a more diverse and representative sample to better understand how different demographic (e.g., gender) and social groups experience safety within the built environment.

7. Conclusions

In conclusion, by integrating eye-tracking systems and street view images, we examined human safety perceptions within the urban environment comprehensively. Several objects in urban environments were identified that draw visual attention and affect human safety perceptions when observing the built environment. We discovered that certain urban infrastructure elements, such as stairways and floors, and public space features, such as flags and chairs attract more human attention when determining safe environments, while the presence of vehicles might be associated with unsafe environments. Open-space features, such as sky, receive less attention despite their high proportions in images. More importantly, by comparing heatmaps generated from eye-tracking systems and those generated using eXplainable AI (XAI) techniques, we evaluate the trustworthiness of these machine-based metrics in measuring safety perceptions. We suggest that the results generated by XGradCAM and EigenCAM are the most aligned with human safety perceptions in a Swedish context. This study demonstrates how integrating emerging technologies including street view images, deep learning, and eye-tracking systems could deepen our knowledge of human–environment relationships, and could inform the future design of safe environments and communities.

CRediT authorship contribution statement

Yuhao Kang: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Junda Chen:** Visualization, Methodology, Formal analysis. **Liu Liu:** Writing – original draft, Methodology, Formal analysis. **Kshitij Sharma:** Writing – review & editing, Data curation, Conceptualization. **Martina Mazzarello:** Data curation, Conceptualization. **Simone Mora:** Writing – review & editing, Project administration, Data curation, Conceptualization. **Fábio Duarte:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Carlo Ratti:** Supervision, Funding acquisition.

Acknowledgments

The authors would like to thank Dr. Paolo Santi at the MIT Senseable City Lab who provided valuable insights about the discussions of the work. The authors would like to acknowledge the City of Helsingborg, and all members of the MIT Senseable City Lab Consortium for their support: FAE Technology, Dubai Future Foundation, Sondotécnica, Arnold Ventures, Sidara, Toyota, A2A, UnipolTech, Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria, Hospital Israelita Albert Einstein, Atlas University, KACST - King Abdulaziz City for Science and Technology, SMART - Singapore-MIT Alliance for Research and Technology, KAIST Center for Advanced Urban Systems, Seoul AI Foundation, AMS Institute, Rio de Janeiro, Laval, and Amsterdam. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders. During the writing of the manuscript, ChatGPT was utilized as a tool solely for proofreading purposes, without contributing any ideas or perspectives to the content of the paper.

Appendix

A.1. Dataset construction

We constructed our sample dataset by selecting 300 street view images among all street view images in Helsingborg, Sweden, to represent varying urban landscapes. To ensure broader geographic representation across the city, we performed a spatially random sampling of street view images. Following this, we manually reviewed and refined the sample by removing images that were either completely black or significantly distorted. We also removed a few images depicting duplicated scenes to enhance the diversity and representativeness of visual contexts in the experiment. Both the global model and the Stockholm model were applied to all images, yielding continuous safety perception ratings on a 1–9 scale.

We then categorized the images into three safety perception levels based on these scores (Lei et al., 2024; Tang, Zeng, Wang, Zhang, & Xu, 2024): *high*, *medium*, and *low*. Adopting such a ternary classification rather than directly using continuous safety perception scores (1–9) was intended to mitigate the following two potential issues: (1) subtle differences in a continuous scoring scale may not be perceptually significant to participants; and (2) to reduce the influences on individual differences in eye-tracking experiments (Valuch, Pflüger, Wallner, Laeng, & Ansorge, 2015). Each category contains 100 street-view images to be utilized in our further experiments. Specifically, images classified within the top 20% according to both scores were designated as having *high safety* perceptions, whereas those in the bottom 20% were assigned as having *low safety* perceptions. Images that fell between the 40% and 60% percentiles were considered to have *medium safety* perceptions. We excluded images in the 20%–40% and 60%–80% ranges to avoid ambiguous stimuli and to maintain clear perceptual contrasts among categories. Upon creating these categories, we randomly selected 100 images from each safety perception category and compiled a balanced dataset of 300 images. It should be noted that the three-category classification of safety perception scores (high, medium, low) was applied solely as a pre-processing step to construct a balanced sample dataset for subsequent analysis. Such a categorization does not imply a direct or absolute relationship between the assigned score and the safety perceptions evoked in later user studies. An image with a high safety perception score does not necessarily elicit strong feelings of safety from participants. This demonstrates the importance of keeping a human-in-the-loop approach, where human behavior responses are important rather than relying solely on algorithmic predictions. The number of selected images was determined by our preliminary estimates of the number of participants in our study, aiming to strike a balance between the frequency of image occurrences and the number of participants. This consideration ensures that most images were compared with other images over five times, thereby facilitating more reliable estimations of human safety perceptions. This dataset will be further employed in the subsequent survey to collect human safety perceptions with eye-tracking systems in Section 3.2.

A.2. Image similarity metrics

We offer more technical details about the L2 loss and LPIPS scores leveraged for measuring image similarities in Section 5.2.

The L2 loss is computed as the Euclidean distance between the corresponding pixel values $p(i, j)$ of two images m and n.

$$L2\ \text{loss} = \sqrt{\sum_{i=1}^{h}\sum_{j=1}^{w}(p_m(i,j) - p_n(i,j))^2} \tag{3}$$

While the L2 loss has been commonly used as a basic metric for evaluating image similarities, it may not fully represent the perceived similarity because it is sensitive to individual pixel matches. Furthermore, it is challenging to understand the meanings of such error representation metrics. Thus, we utilized the LPIPS score, a validated metric for measuring perceptual similarity between image pairs in prior studies, to evaluate the similarity between human attention heatmaps and XAI-based heatmaps (Jang et al., 2024; Zhang, Isola et al., 2018). LPIPS quantifies the Euclidean distance between feature vectors of images extracted from a pretrained deep convolutional neural network, like AlexNet (Krizhevsky, Sutskever, & Hinton, 2012). We also applied this network as a feature extractor to calculate LPIPS scores for comparing two images. By combining these two measures, we measured the image similarity between the human attention and XAI-based heatmap images at the scene level. Results help identify the XAI model that best matches human visual attention.

Data availability

Data will be made available on request.

References

Abubakar, I. R., & Aina, Y. A. (2019). The prospects and challenges of developing more inclusive, safe, resilient and sustainable cities in Nigeria. *Land Use Policy*, *87*, Article 104105.

Appleton, J. (1975). *The experience of landscape*. Wiley.

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, *215*, Article 104217.

Brymer, E., Crabtree, J., & King, R. (2021). Exploring perceptions of how nature recreation benefits mental wellbeing: a qualitative enquiry. *Annals of Leisure Research*, *24*(3), 394–413.

Ceccato, V. (2013). *Moving safely: crime and perceived safety in Stockholm's subway stations*. Lexington books.

Ceccato, V., Kang, Y., Abraham, J., Näsman, P., Duarte, F., Gao, S., et al. (2025). What makes a place safe? Assessing AI-generated safety perception scores using Stockholm's street view images. *The British Journal of Criminology*, azaf017.

Ceccato, V., & Newton, A. (2015). Aim, scope, conceptual framework and definitions. *Safety and Security in Transit Environments: An Interdisciplinary Approach*, 3–22.

Dong, L., Jiang, H., Li, W., Qiu, B., Wang, H., & Qiu, W. (2023). Assessing impacts of objective features and subjective perceptions of street environment on running amount: A case study of Boston. *Landscape and Urban Planning*, *235*, Article 104756.

Dong, W., Liao, H., Roth, R. E., & Wang, S. (2014). Eye tracking to explore the potential of enhanced imagery basemaps in web mapping. *The Cartographic Journal*, *51*(4), 313–329.

Doran, B. J., & Lees, B. G. (2005). Investigating the spatiotemporal links between disorder, crime, and the fear of crime. *The Professional Geographer*, *57*(1), 1–12.

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part i 14* (pp. 196–212). Springer.

Fisher, B. S., & Nasar, J. L. (1992). Fear of crime in relation to three exterior site features: Prospect, refuge, and escape. *Environment and Behavior*, *24*(1), 35–65.

Gobster, P. H., & Westphal, L. M. (2004). The human dimensions of urban greenways: planning for recreation and related experiences. *Landscape and Urban Planning*, *68*(2–3), 147–165.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645.

Hamim, O. F., & Ukkusuri, S. V. (2024). Towards safer streets: A framework for unveiling pedestrians' perceived road safety using street view imagery. *Accident Analysis and Prevention*, *195*, Article 107400.

He, B., Dong, W., Liao, H., Ying, Q., Shi, B., Liu, J., et al. (2023). A geospatial image based eye movement dataset for cartography and GIS. *Cartography and Geographic Information Science*, *50*(1), 96–111.

He, B., Qin, T., Shi, B., & Dong, W. (2024). How do human detect targets of remote sensing images with visual attention? *International Journal of Applied Earth Observation and Geoinformation*, *132*, Article 104044.

Hei, Q., Yang, T., Dong, W., He, B., & Han, D. (2025). Leveraging psychology and neuroscience for geospatial cognition research. *Annals of GIS*, 1–13.

Hollander, J. B., Purdy, A., Wiley, A., Foster, V., Jacob, R. J., Taylor, H. A., et al. (2018). Seeing the city: Using eye-tracking technology to explore cognitive responses to the built environment. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*.

Ito, K., Kang, Y., Zhang, Y., Zhang, F., & Biljecki, F. (2024). Understanding urban perception with visual data: A systematic review. *Cities*, *152*, Article 105169.

Jacobs, J. (1961). *The death and life of great American cities*. New York: Random House.

Jang, K. M., Chen, J., Kang, Y., Kim, J., Lee, J., & Duarte, F. (2023). Understanding place identity with generative AI. arXiv preprint arXiv:2306.04662.

Jang, K. M., Chen, J., Kang, Y., Kim, J., Lee, J., Duarte, F., et al. (2024). Place identity: a generative AI's perspective. *Humanities and Social Sciences Communications*, *11*(1), 1–16.

Jeffery, C. R. (1971). Crime prevention through environmental design. *American Behavioral Scientist*, *14*(4), 598.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Kang, Y. (2023). *Understanding human perception of place with geospatial data science* (Unpublished doctoral dissertation), The University of Wisconsin-Madison.

Kang, Y. (2025). Human-centered geospatial data science. arXiv preprint arXiv:2501.05595.

Kang, Y., Abraham, J., Ceccato, V., Duarte, F., Gao, S., Ljungqvist, L., et al. (2023). Assessing differences in safety perceptions using GeoAI and survey across neighbourhoods in Stockholm, Sweden. *Landscape and Urban Planning*, *236*, Article 104768.

Kang, Y., Gao, S., & Roth, R. E. (2024). Artificial intelligence studies in cartography: a review and synthesis of methods, applications, and ethics. *Cartography and Geographic Information Science*, 1–32.

Kang, Y., Zhang, F., Gao, S., Lin, H., & Liu, Y. (2020). A review of urban physical environment sensing using street view imagery in public health studies. *Annals of GIS*, *26*(3), 261–275.

Kiefer, P., Giannopoulos, I., Raubal, M., & Duchowski, A. (2017). Eye tracking for spatial research: Cognition, computation, challenges. *Spatial Cognition & Computation*, *17*(1–2), 1–19.

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

Larkin, A., Gu, X., Chen, L., & Hystad, P. (2021). Predicting perceptions of the built environment using GIS, satellite and street view image approaches. *Landscape and Urban Planning*, *216*, Article 104257.

Lavie, N., Hirst, A., De Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, *133*(3), 339.

Lei, Y., Zhou, H., Xue, L., Yuan, L., Liu, Y., Wang, M., et al. (2024). Evaluating and comparing human perceptions of streets in two megacities by integrating street-view images, deep learning, and space syntax. *Buildings*, *14*(6), 1847.

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.

Ma, H., Zhang, Y., Cui, Q., Kwan, M.-P., & Cai, J. (2025). Perceived distance to greenery affects psychological restoration. *Environment and Planning B: Urban Analytics and City Science*, Article 23998083251362610.

Moreno-Vera, F. (2021). Understanding safety based on urban perception. In *International conference on intelligent computing* (pp. 54–64). Springer.

Nasar, J. L., & Jones, K. M. (1997). Landscapes of fear and stress. *Environment and Behavior*, *29*(3), 291–323.

Newman, O. (1973). *Defensible space: Crime prevention through urban design*. Collier Books New York.

O'Brien, D. T., Farrell, C., & Welsh, B. C. (2019). Looking through broken windows: The impact of neighborhood disorder on aggression and fear of crime is an artifact of research design. *Annual Review of Criminology*, *2*, 53–71.

Ogawa, Y., Oki, T., Zhao, C., Sekimoto, Y., & Shimizu, C. (2024). Evaluating the subjective perceptions of streetscapes using street-view images. *Landscape and Urban Planning*, *247*, Article 105073.

Qin, T., Dong, W., & Huang, H. (2023). Perceptions of space and time of public transport travel associated with human brain activities: a case study of bus travel in Beijing. *Computers, Environment and Urban Systems*, *99*, Article 101919.

Raco, M. (2007). Securing sustainable communities: Citizenship, safety and sustainability in the new urban planning. *European Urban and Regional Studies*, *14*(4), 305–320.

Rahm, J., Sternudd, C., & Johansson, M. (2021). "In the evening, I don't walk in the park": The interplay between street lighting and greenery in perceived safety. *Urban Design International*, *26*, 42–52.

Raiko, T., Valpola, H., & LeCun, Y. (2012). Deep learning made easier by linear transformations in perceptrons. In *Artificial intelligence and statistics* (pp. 924–932). PMLR.

Ramanujam, P. (2006). *Prospect-refuge theory revisited: A search for safety in dynamic public spaces with a reference to design*. The University of Texas at Arlington.

Ramírez, T., Hurtubia, R., Lobel, H., & Rossetti, T. (2021). Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety. *Landscape and Urban Planning*, *208*, Article 104002.

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12179–12188).

Raschke, M., Blascheck, T., & Burch, M. (2014). Visual analysis of eye tracking data. *Handbook of Human Centric Visualization*, 391–409.

Rodriguez-Spahia, D., & Barberet, R. (2020). Inclusive and safe cities for the future: A criminological analysis. In *The emerald handbook of crime, justice and sustainable development* (pp. 223–241). Emerald Publishing Limited.

Sangers, R., van Gemert, J., & van Cranenburgh, S. (2022). Explainability of deep learning models for urban space perception. arXiv preprint arXiv:2208.13555.

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).

Shaw, S.-L., & Sui, D. (2021). Understanding the new human dynamics in smart spaces and places: Toward a splatial framework. In *Smart spaces and places* (pp. 7–16). Routledge.

Stamps III, A. E. (2005). Enclosure and safety in urbanscapes. *Environment and Behavior*, *37*(1), 102–133.

Tabrizian, P., Baran, P. K., Smith, W. R., & Meentemeyer, R. K. (2018). Exploring perceived restoration potential of urban green enclosure through immersive virtual environments. *Journal of Environmental Psychology*, *55*, 99–109.

Tang, F., Zeng, P., Wang, L., Zhang, L., & Xu, W. (2024). Urban perception evaluation and street refinement governance supported by street view visual elements analysis. *Remote Sensing*, *16*(19), 3661.

Un-Habitat (2012). *Enhancing urban safety and security: Global report on human settlements 2007*. Routledge.

Uttley, J., Simpson, J., & Qasem, H. (2018). Eye-tracking in the real world: Insights about the urban environment. In *Handbook of research on perception-driven approaches to urban assessment and design* (pp. 368–396). IGI Global.

Valuch, C., Pflüger, L. S., Wallner, B., Laeng, B., & Ansorge, U. (2015). Using eye tracking to test for individual differences in attention to attractive faces. *Frontiers in Psychology*, *6*, 42.

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., et al. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(2), 7–30.

Wang, R., Yuan, Y., Liu, Y., Zhang, J., Liu, P., Lu, Y., et al. (2019). Using street view data and machine learning to assess how perception of neighborhood safety influences urban residents' mental health. *Health & Place*, *59*, Article 102186.

Wilson James, Q., & Kelling George, L. (1982). Broken windows: The police and neighborhood safety. *Atlantic Monthly*, *249*(3), 29–38.

Yang, N., Deng, Z., Hu, F., Chao, Y., Wan, L., Guan, Q., et al. (2024). Urban perception by using eye movement data on street view images. *Transactions in GIS*, *28*(5), 1021–1042.

Yang, T., Qin, T., Zhang, J., Dong, Z., Wu, Y., Wan, X., et al. (2025). Neurocognitive geography: exploring the nexus between geographic environments, the human brain, and behavior. *Science Bulletin*, S2095–9273.

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.

Ye, X., Du, J., Li, X., Shaw, S.-L., Fu, Y., Dong, X., et al. (2025). Human-centered geoai foundation models: where geoai meets human dynamics. *Urban Informatics*, *4*(1), 2.

Zhang, F., Fan, Z., Kang, Y., Hu, Y., & Ratti, C. (2021). "Perception bias": Deciphering a mismatch between urban crime and perception of safety. *Landscape and Urban Planning*, *207*, Article 104003.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586–595).

Zhang, W., & Liu, H. (2017). Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications. *IEEE Transactions on Image Processing*, *26*(5), 2424–2437.

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.

Zhao, B., Feng, J., Sun, Y., Chang, X., & Shaw, S.-L. (2024). Neural sensing: Toward a new approach to understanding emotional responses to place. *Transactions in GIS*, *28*(7), 2463–2475.

Zhou, H., Wang, J., Wilson, K., Widener, M., Wu, D. Y., & Xu, E. (2025). Using street view imagery and localized crowdsourcing survey to model perceived safety of the visual built environment by gender. *International Journal of Applied Earth Observation and Geoinformation*, *139*, Article 104421.

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).