This is the Accepted Manuscript version of an article published by Elsevier in the journal *Landscape and Urban Planning* in 2025, which is available at: https://doi.org/10.1016/j.landurbplan.2025.105368
Cite as: Gu Y, Quintana M, Liang X, Ito K, Yap W, Biljecki F (2025): Designing effective image-based surveys for urban visual perception. *Landscape and Urban Planning*, 260: 105368.

Designing Effective Image-based Surveys for Urban Visual Perception

Youlong Gu[a,b], Matias Quintana[b], Xiucheng Liang[a], Koichi Ito[a], Winston Yap[a], Filip Biljecki[a,c,*]

[a]*Department of Architecture, National University of Singapore, 4 Architecture Dr, 117566, Singapore*
[b]*Future Cities Lab Global, Singapore-ETH Centre, CREATE campus, 1 Create Way, #06-01 CREATE Tower, 138602, Singapore*
[c]*Department of Real Estate, National University of Singapore, 15 Kent Ridge Drive, 119245, Singapore*

Abstract

Urban visual perception is important for the human experience in cities, shaped by intertwined characteristics of urban landscapes. By quantifying and explaining these perceptual experiences, researchers can gain insights into human preferences and support decision-making in planning and design. However, past studies have shown inconsistencies in survey design and ambiguities in reporting, leading to concerns about the reliability and reproducibility of results. This study proposes the first comprehensive framework to guide image-based survey design for capturing perceptions of outdoor urban environments across different scenarios, addressing the lack of methodological standardization in current research. We reviewed existing surveys to identify key parameters, conducted comprehensive between-subject and within-subject surveys, and performed statistical analyses to determine best practices for survey design across different contexts. Aiming to set a potential community standard, our study doubles as a blueprint for a reporting protocol for survey designs. Based on the results, we recommend: (1) meeting a minimum of 12 and 22 ratings per image for Likert Scale and Pairwise Comparison studies to reach survey reliability, respectively, and reporting these alongside other survey design parameters to enhance transparency and reproducibility; and (2) when resource allows larger experiments, adopt a ranking method such as Pairwise Comparison to achieve firmer rating results; and (3) using perspective (non-panoramic) images more frequently, as they exhibit comparable overall scores to panoramic images (R mostly > 0.7), while being more widely available via crowdsourced sources, supporting their use in large-scale visual perception research.

Keywords: urban perception, survey parameters, human participants, built environment, street view imagery

*Corresponding author. Email: filip@nus.edu.sg

1. Introduction

Visual perception refers to the process by which humans interpret visual information from the external environment, forming an understanding of surroundings (Gibson, 2014; Snowden et al., 2012; Cornsweet, 2012; Wade and Swanston, 2012). With the accelerating urbanization, more studies are focusing on the impact of urban environments on human visual perception, evaluating people's preferences and needs in them through data such as images and videos (Ito et al., 2024; Dubey et al., 2016; Zhang et al., 2018). Specifically, understanding human subjective cognition helps researchers better understand urban environments and leads to more holistic and human-oriented evaluations of aspects such as environmental safety (Liu et al., 2025; Wang et al., 2022; von Stülpnagel and Binnig, 2022), physical and mental health (Wu et al., 2024; Chen et al., 2022; Zhang et al., 2021), and aesthetic preference (Wang et al., 2019b; Gao et al., 2019), providing crucial information for decision-making in urban planning.

The evaluation of urban visual perception is usually conducted quantitatively. To study it, researchers have traditionally relied on methods such as field observations, questionnaires, and interviews (McGinn et al., 2007; Clay and Smidt, 2004; Oku and Fukamachi, 2006; Brush et al., 2000). These methods are typically straightforward but are time-consuming, costly, and limited in sampling, making them suitable only for small-scale studies. In recent years, with the advancement in computer vision and the proliferation of image sources such as Street View Imagery (SVI) services and their accessibility, large-scale online surveys based on visual data became a norm (Ito et al., 2024; Biljecki and Ito, 2021; Anguelov et al., 2010; Yan et al., 2020). Many such studies show no sign of relenting, with researchers using automated auditing workflows and ever-increasing visual data sources to conduct large-scale urban perception surveys (Kelly et al., 2013; Rundle et al., 2011).

Often constrained by various institutional and budgetary considerations, researchers typically design such surveys independently. As a result, the criteria and parameters they adopt may appear arbitrary or, at best, reflect weak coordination within the international scientific community, and in many cases, the specifics remain rather vague. Key reasons for that are the lack of universally standardized guidelines and a foundational study that addresses this topic.

Because of this gap, researchers decide several survey parameters separately, such as choices about the types of images used, the format of survey questions, and the sample size (or number of images rated). For example, some studies use panoramas to assess urban walkability (Huang et al., 2023), while others utilize perspective images with limited viewpoints to analyze the same aspect (Gong et al., 2023). Some studies evaluate each image fewer than 10 times (Kang et al., 2023; Luo et al., 2022b), while others conduct more than 100 assessments per image (Ogawa et al., 2024). These decisions introduce flexibility and variations, leading to potentially inconsistent results. Moreover, the aforementioned parameters themselves are often not clearly defined or remain ambiguous in publications. Without standardized guidelines, such variations and ambiguities may yield conclusions that may not be directly comparable or, in some cases, may not be valid.

We call for bridging these gaps by giving attention to the foundation of this research line and by providing a set of guidelines supported by statistics and comprehensive multi-dimensional ex-

periments, which would drive future surveys. Given the diverse practices, survey designs, and parameter choices in studies, we seek to answer the following research questions (RQ), accompanied by multiple subquestions:

How can we develop a set of standardized approaches and protocols for conducting and reporting robust and effective visual perception surveys in urban studies?

- RQ1: How should the consistency of responses from different raters be evaluated?
- RQ2: To what extent do different survey parameters affect the results?
- RQ3: Is there a minimum sample size setting to ensure robustness?

To answer them, we have established a comprehensive framework backed by statistics to ensure sound approaches and standardization of image-based perception surveys. We identified common parameters in visual perception surveys, designed and conducted a comprehensive survey based on these parameters, and performed data analysis using a statistical framework that encompasses consistency, discrepancy, and robustness. This framework explores best practices for conducting effective surveys and the advantages and disadvantages of parameter choices in different circumstances. Further, it helps communicate survey designs effectively. The framework aims to help researchers in related fields reach a consensus on survey design and assist those unfamiliar with it in understanding common variables and their suitability. When doing so, we have considered a wide range of strands in urban planning and design for which perception studies with imagery are employed, from safety and aesthetics to comfort and walkability, intending to render this work as application-agnostic and as widely applicable as possible.

2. Literature review

2.1. Visual perception survey

In recent years, the development of computer vision, along with the emergence of numerous image data sources, has led to the gradual replacement of traditional low-throughput surveys and interviews with efficient large-scale visual perception surveys (Ito et al., 2024; McGinn et al., 2007; Clay and Smidt, 2004; Oku and Fukamachi, 2006; Brush et al., 2000). These surveys typically use visual interfaces to display visuals, and raters assess them based on their intuitive feelings, conducting various types of evaluations, depending on the goal of the survey. Driven by technological advancements and shifting research interests, studies on urban visual perception have proliferated and cover a wide spectrum of topics, such as perception of street safety, landscape preferences, and the quality of the built environment (Ito et al., 2024).

While studies tend to collect ratings on their own, there are instances utilizing existing ratings from open datasets. The Place Pulse datasets serve as rare and marquee examples of open visual perception data, having collected scores from over 80,000 individuals worldwide since the early 2010s (Salesses et al., 2013; Dubey et al., 2016). These open datasets have enabled many

researchers to rely on the existing local annotations within these datasets and combine them with other multi-source data to evaluate the distribution of human preferences on a larger scale, without having to collect data on their own. Having a set of six dimensions (e.g., scoring safety, beauty), Place Pulse has shaped the contours of this field, as the same dimensions have been adopted also by many research groups conducting surveys on their own, independently of this dataset (to be discussed later in detail). However, as other large-scale crowdsourced data, they lack comprehensive evaluation of data quality. The heterogeneity (e.g., varying numbers of scores per image) raises concerns about their reliability, which propagates to the different degrees of robustness, reliability, and validity of the studies utilizing them.

Besides the aforementioned data, most researchers resort to collecting data independently for the lack of high-quality open data instances and significant differences in research objectives. In these studies, the choice of survey parameters, such as number of participants, varies significantly and is driven by many factors, such as budget, available time and data, scope, and focus of the study. For example, studies use various SVI types: panoramic images (Zhao et al., 2023; Huang et al., 2023; Chen and Biljecki, 2023), perspective images (Yao et al., 2019; Navarrete-Hernandez et al., 2023; Qiu et al., 2023; Wang et al., 2019a), and panoramic mosaic images (Meir and Oron-Gilad, 2020; Ma et al., 2023). In terms of scoring methods, most studies use the Likert Scale (Lis et al., 2022; Cui et al., 2023), but some are increasingly adopting Pairwise Comparison for ranking scoring. Regarding rater settings (Kang et al., 2023; Ye et al., 2019; Ramírez et al., 2021), there is considerable variation in terms of the rater backgrounds, which can include industry experts, student groups, the general public, or their concoction. Due to the varying number of raters and the images being evaluated, the number of assessments per image also differs significantly. This can range from fewer than 10 evaluations (Kruse et al., 2021; Ito and Biljecki, 2021; Luo et al., 2022b), to between 10 and 100 (Qiu et al., 2023; Luo and Jiang, 2022; Ma et al., 2023; Navarrete-Hernandez et al., 2023; Zhou et al., 2025), or even exceed 100 assessments per image (Ogawa et al., 2024). In this context, researchers face the challenge of determining the most effective survey practices to evaluate rater consistency and decide on different input parameters (such as image types, scoring methods, and raters). These factors are crucial for the accuracy and reliability of research results and require further exploration and optimization.

Overall, current urban visual perception research presents structural variability due to numerous both simple and complex factors, posing significant challenges to researchers. They face the dilemma of identifying the most effective survey methods to evaluate the consistency of the rating and find the appropriate survey parameters without a foundational set of guidelines grounded in statistics. There is a need to explore and verify the impact of different survey parameter settings on perception results to ensure the accuracy and reliability of studies, which we aspire to bridge in this paper.

2.2. Effective survey design with statistics

On a general scope, surveys are an essential channel for obtaining and analyzing information from a sample of individuals (Groves et al., 2011; Rea and Parker, 2014; Fink, 2003; Krosnick, 1999). Conducting a comprehensive survey is usually a complex process that includes survey

design, data collection, data editing, and data analysis, which requires careful consideration and decision-making. A well-designed survey is fundamental to ensuring data quality, representativeness, and the credibility of results (Bethlehem, 2009; Groves et al., 2011; Fink, 2003). Therefore, guiding effective survey design has long been important.

An effective survey, when designed using appropriate methodologies, can collect accurate data within the constraints of limited resources, thereby fulfilling the needs of downstream tasks such as statistical analysis (Taherdoost, 2016; Kasunic, 2005). Survey guidelines provide recommendations for effective survey design by identifying key aspects such as question formulation, instrument design, and sample selection across different contexts. For instance, Rea and Parker (2014) and Fink (2003) provided comprehensive survey guides that span from question development and sample estimation to survey implementation and data analysis optimization. To enhance effectiveness, some guidelines and exploratory analyses specifically focus on parameters such as survey reliability and validity (Aithal and Aithal, 2020; Gadermann et al., 2012; Hallgren, 2012), sample selection strategies (Dell et al., 2002; Eng, 2003; Lenth, 2001) and participation optimization (Salminen et al., 2025).

Beyond general methodologies, specific fields have developed detailed guidelines tailored to their unique practices for guiding survey design in particular use cases. For example, in the medical field, Bennett et al. (2011) used literature reviews to emphasize the importance of survey reporting, and Artino et al. (2018) provided a reporting protocol of survey rationale, pre-testing, and implementation. In the field of biological monitoring, the focus has been on survey design options, cost functions, and the construction of models, highlighting the significance of study planning for effective survey design (Reynolds et al., 2011). In software engineering, past survey parameters and related issues, such as target audience, survey tools, data analysis, and reporting, have been identified, with optimization recommendations based on these findings (Molléri et al., 2016). In psychology, expert consensus methods have been employed to develop quality checklists for survey research (Protogerou and Hagger, 2020), and in another study, five guidelines were proposed to create a more gender-inclusive science, aiming to avoid validity issues stemming from gender measurement errors (Cameron and Stinson, 2019).

Despite the widespread application of survey guidelines across various fields, the urban visual perception domain remains underexplored in terms of a unified survey framework. Given the unique practices in this field — such as image types (e.g., Street View Imagery) and application contexts (e.g., urban perceived safety) — no comprehensive survey guideline has yet compared common practices or provided concrete recommendations for future surveys. Addressing this gap, our study offers a comprehensive framework for identifying common parameters in urban visual perception research, providing the first reference for the field. Our aim is to guide the research community toward establishing consensus on selecting, designing, and reporting survey parameters in this context.

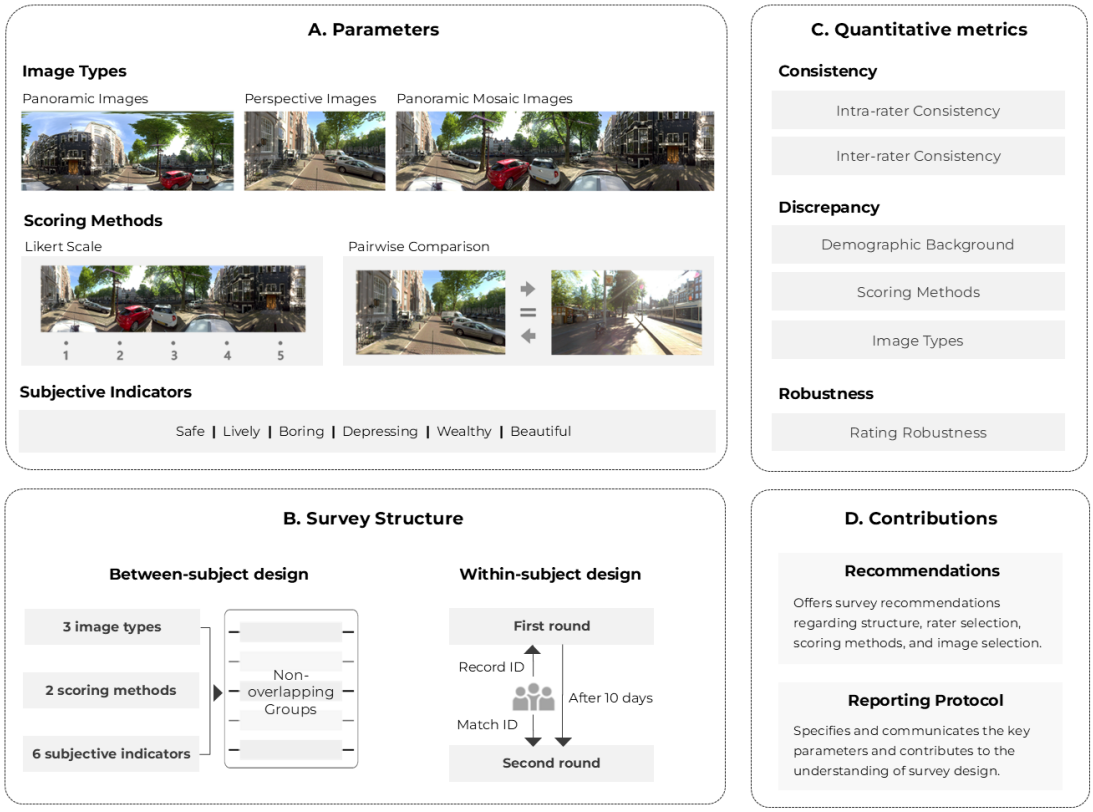

Figure 1: A framework for study design encompassing the most relevant parameters (A), survey structure (B), three quantitative metrics addressing the research questions (C), and the overall contributions (D).

3. Methodology

The research framework is illustrated in Figure 1. Initially, we base our approach on existing practices utilizing outdoor environment image data for surveys, selecting representative parameters (Figure 1A) at three levels: Image Types, Scoring Methods, and Subjective Indicators for comparative analysis. Subsequently, we design a robust survey structure through between-subject and within-subject design (Figure 1B) to assess inter-rater and intra-rater agreement, respectively. Finally, we conduct quantitative analyses based on three metrics — Consistency, Discrepancy, and Robustness — to provide recommendations and establish protocols (Figure 1C).

3.1. Parameters

3.1.1. Image types

According to Ito et al. (2024), images constitute nearly all visual data used in this domain. We have chosen Street View Imagery (SVI) because it has gained significant momentum thanks

to services such as Google Street View and Baidu Maps (Anguelov et al., 2010; Biljecki and Ito, 2021). Moreover, SVI encompasses a variety of image types, making it an excellent proxy for other image types, such as non-geo-tagged photos. We have selected the three most common image types to compare the influence of this factor.

- Panoramic Images

  These are a form of wide-angle representation, commonly provided by commercial street view services, and are created by stitching multiple perspective images. Panoramic images contain comprehensive scene information but often exhibit varying degrees of perspective distortion due to their projection, which leads to local magnification or reduction of certain elements depending on their distance from the center of perspective.

- Perspective Images

  Such images represent scenes as they appear to humans, and are commonly provided by non-geo-tagged photo sources and crowdsourced SVI platforms (e.g., Mapillary, KartaView) (Hou et al., 2024; Wang et al., 2024). These images can be captured using different devices or derived from panoramic images by slicing them. Perspective images contain scenes from specific angles, eliminating the distortion effect, but losing some edge pixels (Beaucamp et al., 2022), resulting in element proportions that are closer to the real-world observation by the human eye. In the case of SVI, perspective images with views along the direction of a road or walking path provide results closer to the overall scene (Biljecki et al., 2023), which we select in this study.

- Panoramic Mosaic Images

  These are created by stitching multiple nearly distortion-free perspective images to form a complete 360 view. The field of view of a single image is similarly limited to that of perspective images, but due to the nature of stitching them, the overall coverage is between that of single perspective images and that of panoramic ones. In this study, panoramic mosaic images are set to four directions of SVI (0, 90, 180, 270), with an aspect ratio = 1:1. To emphasize the differences in image presentation compared to panoramic images while maintaining the original elements, the stitching order of the four SVIs is adjusted to (180, 270, 0, 90).

3.1.2. Scoring methods

Following the preparation of images, multiple, considerably different, scoring methods are employed in surveys, often without much reasoning about the choice. The most common scoring methods are usually divided into rating scoring and ranking scoring (Ak et al., 2021). In these two common practices, we have selected one method in each as a representative.

- Likert Scale

This scoring method, proposed by sociologist Rensis Likert in the 1930s (Likert, 1932; Wu and Leung, 2017), is widely applied in social science (Jebb et al., 2021). In visual perception surveys, the Likert Scale is commonly employed, where raters select from options ranging from 'strongly disagree' to 'strongly agree' to reflect their attitudes on images. This study uses a five-point Likert scale, the minimum recommended for treating option ordinal scores as interval data during mathematical and statistical calculations (Harpe, 2015).

The scoring calculation method in the Likert Scale is relatively straightforward. Once the minimum number of options is met and can be approximated as interval data, the options are digitized from low to high (e.g., from 1 to 5). The score is then calculated as the mean of all rating responses, and it is simply termed **Mean Score**.

- Pairwise Comparison

  This ranking method, proposed by psychologist, L. L. Thurstone (Thurstone, 1974), has been emphasized in recent years in online visual surveys. It involves presenting pairs of images to raters, who choose the one with a higher intensity of a given attribute or select the 'equal' option if they perceive the intensity equally. The results of such comparisons are inherently relative and hold meaning only within the specific rater pool and dataset from which they were derived.

  According to common practice, Pairwise Comparison scoring calculation involves two common methods: strength of schedule (SOS) and ranking system algorithms.

  The strength of schedule (SOS) method was first introduced to the urban perception domain by Salesses et al. (2013). The win (W) and loss (L) ratios of image i with respect to question u are defined as follows:

$$W_{i,u} = \frac{w_{i,u}}{w_{i,u} + l_{i,u} + t_{i,u}}, \quad L_{i,u} = \frac{l_{i,u}}{w_{i,u} + l_{i,u} + t_{i,u}} \tag{1}$$

  where w is the number of times an image was selected over its paired image, l is the number of times an image was not chosen over its paired image, and t is the number of times when an image was chosen as equal to its paired image. Using this, we define the score for each image i and question u as:

$$Q_{i,u} = \frac{10}{3}\left(W_{i,u} + \frac{1}{n_i^w}\sum_{j_1=1}^{n_i^w} W_{j_1 u} - \frac{1}{n_i^l}\sum_{j_2=1}^{n_i^l} L_{j_2 u} + 1\right) \tag{2}$$

  where n_i^w is the total number of images i was preferred over, n_i^l is the total number of images i was not preferred over, the first sum extends over j_1, the images that image i was preferred over, and the second sum extends over j_2, the images that were preferred over i. This method's score is referred to as the **Q Score**.

  Common ranking system algorithms include the Elo ranking system and the Trueskill rank-

ing system. These algorithms share the characteristic of inferring each entity's score and corresponding ranking from Pairwise Comparison. We follow the Trueskill ranking system as the representative method due to its faster convergence, broader applicability, and more stable evaluations compared to the other common ranking methods such as the Elo system, with the resulting score termed the **Trueskill Score** (Herbrich et al., 2006).

3.1.3. Subjective indicators

Since Dubey et al. (2016) first proposed six general indicators of urban perception — safe, lively, boring, depressing, wealthy, and beautiful — in Place Pulse, these indicators or their subsets have been consistently used in various studies related to the distribution and preferences of urban perception indicators (Cui et al., 2023; Kang et al., 2023; Meir and Oron-Gilad, 2020; Hidayati et al., 2020). While other aspects of urban visual perception studies, such as landscape-centric (Shayestefar et al., 2022; Suppakittpaisarn et al., 2019; Hung and Chang, 2022; Wang et al., 2019b) or pedestrian/cyclist-centric indicators (Ito and Biljecki, 2021; Gong et al., 2023; Kang et al., 2013; Zeng et al., 2024), have been explored, these tend to be highly use-case-specific. In contrast, the six traditional indicators are characterized by their general applicability, ease of interpretation, and suitability for representation through SVIs. To enable comparative analysis with prior research and to establish a consensus for future studies, this research adopts the same six indicators for analysis.

3.2. Survey structure

To assess inter-rater consistency and intra-rater consistency, the survey is structured using a between-subject design for the former and a within-subject design for the latter.

In the first, questions in the dataset, each featuring one of three image types and one of two scoring methods, were selected and evenly distributed across different non-overlapping groups. Within the framework formed by image types and scoring methods, the images and six indicators in each group were fixed. However, the display order for each rater was randomized to eliminate potential bias caused by the sequence. Under this setup, it became possible to compare the overall inter-rater consistency across different groups and evaluate the individual consistency of responses to the same questions within the same group. To ensure that the scores from different scoring methods exceed the minimum requirement and converge stably, we set the scoring frequency for single images using the Likert Scale and Pairwise Comparison to more than double the common use cases (approximately 60 and 90 ratings, respectively).

In the within-subject design, the survey is divided into two rounds. After completing the first round, the system records the question type and sequence, including images, scoring method and indicators assigned to each rater. Ten days later, the raters are invited to complete the second survey, which is identical to the first, but they do not have access to their previous responses. By comparing the responses of the same participants across the two rounds, we can evaluate their intra-rater consistency.

We also collect background information from raters, such as gender, age, nationality, and major (Architecture/Urban-related and non-related majors), to investigate their impact on the results. Information such as survey duration is recorded as well to support the exploratory analysis.

3.3. Quantitative metrics

Tackling the three research questions, we perform a comprehensive analysis of three aspects: Consistency, Discrepancy, and Robustness. This analysis involves evaluating the consistency among raters, analyzing the impact of different survey input parameters on the results and their variability, and determining the minimum sample size that ensures robust outcomes.

3.3.1. Consistency

The objective of this section is to examine the inter-rater and intra-rater consistency to evaluate the reliability of responses under different conditions. Given that our results include both nominal data (from Pairwise Comparison) and ordinal data (from Likert Scale), we cannot apply common methods such as the Intraclass Correlation Coefficient (ICC) or other methods designed for interval variables in a unified way. Instead, we employed Cohen's Kappa and Weighted Kappa methods which are appropriate for handling these data types. The former measures the categorical agreement between two observers while accounting for chance agreement, whereas Weighted Kappa extends this by applying differential weights to discrepancies — with penalties increasing as the difference between responses grows. Notably, quadratic weighting amplifies these differences more significantly than linear weighting. Together, these complementary methods effectively capture the subtle response variations and the impact on overall reliability results.

For the inter-rater consistency check, we only consider the results from the first round of the survey. The survey question set is divided into multiple parts of equal length, and the Weighted Kappa statistic is calculated to evaluate the consistency of responses from different raters to the same question set. To estimate consistency among multiple responses, we calculate the Kappa statistic for each pair of responses and average them to obtain the overall Kappa statistic for the question set. We then calculate the average Kappa statistic for other sets of equal length and compare their standard deviations. Next, we divide the question set into parts of different lengths, calculate their standard deviations individually, and compare them to analyze consistency differences across question sets of varying lengths.

For the intra-rater consistency check, we consider the consistency of responses from the same rater in two rounds of the survey. We then calculate the mean and compare the differences between different Kappa statistics from the two scoring methods.

3.3.2. Discrepancy

The aim of this section is to investigate the influence of survey input factors, including rater background, scoring methods, and image types, on the results. By examining the similarities and differences in various input data, we seek to gain a better understanding and consensus for parameter selection in different scenarios.

First, we examine the impact of different rater backgrounds on the results by focusing on parameters with significant differences, such as gender and major. Using independent samples t-tests, we assess whether these different groups significantly influence overall scores.

In comparing scoring methods, we aim to focus on variables that are applicable to both methods to assess their joint impact on the scoring results. In other words, parameters unique to each

method, such as the number of scales in the Likert scale, are excluded. Consequently, the indicators we focus on are overall score, scoring duration, scoring stability, and the number of minimum ratings required. To achieve this, we compare the association between three types of overall scores (Mean Score, Q Score, and Trueskill Score) using Pearson's correlation. Additionally, we assess the differences in survey duration using a t-test. Scoring stability is evaluated based on the variability of the same method under different parameters. The required number of ratings is examined in detail in the following section.

To contrast the differences based on the image types, we examine the overall score (the final score for a single image) and the raw scoring (the collection of original scores for a single image from different raters) distributions of the images. We calculate Pearson's correlation to determine the relationship between the aggregated overall scores based on image types.

Recognizing that the overall image scores represent average scores and may not reflect the distribution of the raw data, we also use the Mann-Whitney U test, a non-parametric alternative to the t-test, to evaluate differences between two independent samples without assuming normality. Considering the large number of parallel groups in the survey, we aggregate the data by image type and calculate, for each pair of image types, the proportion of parallel groups in which we can reject the null hypothesis (i.e., no difference in scores between different image types) at a significance level of 0.05. This proportion reflects the strength of the differences between the two compared image types.

To further examine the stability of scores across the three image types, we calculate the standard deviation of scores for each image within each group and compare the differences in these standard deviations using a t-test. Additionally, we use kernel density estimation to visualize the distribution of the standard deviations across the three image types.

3.3.3. Robustness

This element primarily investigates the range of scoring frequencies required to achieve the expected robustness of different scoring methods and evaluates the differences in the efficacy of these methods.

For the relationship between the number of raters and the robustness of the score, we implemented a more robust method than that of Salesses et al. (2013) through parallel grouping and repeated calculations. The dataset, which contains a total score count of n for a single image, is divided into non-overlapping image subsets. The length of the subsets is denoted as v, where v ranges from 1 to $n/2$ (since the length of non-overlapping subsets cannot exceed half of the total count). The total dataset is segmented into subsets of length v, and the corresponding scores for each subset are calculated. Finally, the robustness metric is calculated as follows:

$$\text{Robustness}(v) = \left(\frac{\sum_i (S_i^1(v) - \langle S^1(v) \rangle)(S_i^2(v) - \langle S^2(v) \rangle)}{\sigma_1 \sigma_2} \right)^2 \tag{3}$$

where $S_i^1(v)$ and $S_i^2(v)$ are the scores of image i under two different scoring methods, $\langle S^1(v) \rangle$ and $\langle S^2(v) \rangle$ are the mean scores, and σ_1 and σ_2 are the standard deviations of the scores. To minimize

randomness, the process of dividing into non-overlapping subsets is repeated 200 times.

Using this method, we can assess the impact of scoring frequency on score robustness and evaluate several characteristics such as stability and efficacy.

3.4. Study data

3.4.1. Image data and preprocessing

This study employs *Global Streetscapes*, a large open SVI dataset (Hou et al., 2024), which sources imagery from the Mapillary and KartaView platforms from all over the world and constructs semantic indices of images using rich labels (including image type, collection method, quality, location, etc.). Prioritizing panoramic image data availability for its transformability, we selected three stylistically distinct cities among those that have the best data availability: Lima (Peru), Amsterdam (Netherlands), and Athens (Greece). We selected high-quality daytime images captured in clear weather, free of glare or reflections, taken while cycling or driving, and excluding winter seasons (Hou and Biljecki, 2022).

Clustering was also done using segmentation data, resulting in four clusters that effectively represent images dominated by buildings, vegetation, roads, and a mix of these elements in similar proportions. Ten locations were selected in each city, and 2-3 images were randomly drawn from each cluster within the city, yielding 30 panoramic images from these cities. Considering the three image types, the dataset includes a total of 90 images.

Through the aforementioned approach, we ensured both the quality consistency and diversity of images in the dataset. By combining this image data with the estimated sample size from the survey design, we can ensure the reliability and validity of the statistical analysis. With 90 images, two scoring methods, and six subjective indicators, this results in a total of 1,080 combinations.

3.4.2. Survey data

The survey was reviewed and approved by the Institutional Review Board (IRB) of our organization. The questionnaire link was distributed through the university's student work system, and participants — full-time university students — who completed both rounds received cash compensation. Given resource constraints, we selected this specific group because they represent a younger, high-mobility urban demographic, which is particularly relevant for digital surveys in urban studies. Previous research has confirmed a strong correlation between students and the general public's perceptions of the environment (Yao et al., 2012). Furthermore, this choice aligns with existing research that often sources participants from campuses or similarly homogeneous groups (Cao et al., 2025; Lu and Chen, 2024; Qiu et al., 2023; Ye et al., 2019; Kawshalya et al., 2022; Tang and Long, 2019; Li et al., 2022; Gong et al., 2023; Ma et al., 2023). Each rater was required to complete 48 questions in an online questionnaire. Considering the number of images and the required scoring frequency, we estimated that more than 300 participants in the first-round survey would be sufficient to meet the requirements. The first round of the survey was launched in April 2024, and continued until the required number of responses was obtained. Including the second round, the entire process spanned approximately 40 days, spanning several nationalities, and ethnic and cultural backgrounds.

A total of 309 valid responses were collected in the first round, and 281 valid responses in the second round, i.e., achieving a retention rate higher than 90%. In the Likert Scale method, each image received 63 ratings on average, while in the Pairwise Comparison method, each image was rated an average of 92 times. With this comprehensive campaign, we were able to perform diverse analyses to meet our research needs. The proportions of the three image types and six indicators were evenly distributed according to the quota settings in the rating system. Among the demographic information collected, we selected the more balanced sample sizes for the two indicators, gender and major, for analysis. In the first round of data collection, the gender ratio was approximately 1 to 1.2, with 140 males and 169 females. The major ratio was roughly 1 to 5, with 51 students majoring in Architecture or Urban-related fields and 258 students in non-related majors. The gender ratio was approximately 1:1.2 (male to female), and the major ratio was approximately 1:5 (Architecture/Urban-related to non-related majors).

4. Results

4.1. Consistency

We conducted an inter-rater consistency check by dividing the overall question set into equal slices of 10 to 15 groups. For each group, we calculated the Kappa statistics among individuals, evaluated the inter-rater consistency using the average Kappa statistics for the group, and assessed the standard deviations of the consistency among individuals (Table 1). The results indicate that both the average Kappa statistics and their standard deviations exhibit high stability across different slice groupings, with relatively small score fluctuations. This finding suggests that under any given combination of questions, different raters reached a similar level of consensus. Consequently, the random distribution strategy of the survey design proves to be robust, showing consistent data distribution trends across different non-overlapping subsets, which is suitable for subsequent statistical analysis. Specifically, the average Kappa statistic for the Likert Scale is 0.448, slightly lower than 0.589 for Pairwise Comparison. While both methods demonstrated a moderate level of inter-rater consistency, Pairwise Comparison exhibited greater stability and a higher degree of agreement among raters.

We further quantified the intra-rater consistency for the two scoring methods, Figure 2 illustrates three statistics (Cohen's Kappa, Linear Weighted Kappa, and Quadratic Weighted Kappa) and along with a comparison of results after filtering responses based on inter-rater consistency. Cohen's Kappa only considers the consistency of answers without accounting for the degree of scoring differences. For the raw data, this resulted in lower overall consistency under the strictest conditions (Cohen's Kappa Median = 0.516 and 0.581 for the Likert Scale and Pairwise Comparison). After applying linear and quadratic weighting, the consistency of the Likert Scale improved significantly, increasing with higher weights (Linear Weighted Kappa Median = 0.710, Quadratic Weighted Kappa Median = 0.856). Due to the smaller point scale of Pairwise Comparison, the impact on overall statistics remained minimal (Linear Weighted Kappa Median = 0.638, Quadratic Weighted Kappa Median = 0.706). Overall, in Weighted Kappa statistics, the consistency of both scoring methods exceeded 0.6, indicating a substantial agreement level. The Likert Scale method,

Table 1: Kappa statistics and standard deviation of inter-rater consistency.

Method	Group Count	Average Mean Weighted Kappa	Average Standard Deviation
Likert Scale	10	0.457	0.049
	11	0.438	0.062
	12	0.444	0.053
	13	0.485	0.154
	14	0.434	0.075
	15	0.430	0.066
Pairwise Comparison	10	0.582	0.064
	11	0.579	0.054
	12	0.594	0.040
	13	0.592	0.050
	14	0.594	0.040
	15	0.591	0.064

especially with quadratic weighting, showed the highest consistency, suggesting small numerical fluctuations in the Likert Scale have minimal impact on the overall trend.

Using responses within the same non-overlapping group, we compared each rater's responses with those of others to assess individual reliability in the between-subject design. We then filtered out raters with reliability scores below the 25th percentile. After applying this filtering, we calculated the intra-consistency of responses for the two rounds of surveys and compared the results with those from the raw data. The consistency improved across all three Kappa statistics, demonstrating significantly narrower quantile ranges, with all Weighted Kappa statistics exceeding 0.7, indicating a substantial agreement level. This finding demonstrates that the non-overlapping group structure in the between-subject design can effectively identify individual noise and enhance data quality. Consequently, in scenarios where survey resources are limited, and only a single round of data collection is feasible, this approach provides a practical solution for effectively filtering out noisy data.

4.2. Discrepancy

We first examine the impact of different raters' backgrounds on overall scores. Although we collected data on four demographic indicators, we selected only two — gender and major — because these had sufficiently large sample sizes for our analysis. Age was excluded, as most participants were concentrated in the 20-29 age range, and nationality was excluded, as the majority of respondents were Singaporean. By focusing on these two variables, we ensured the validity of our statistical analysis. The kernel density estimate (KDE) distributions and the t-test results of the combined scores for both scoring methods are shown in Figure 3. The results indicate that at an α level of 0.1, significant differences are observed across different demographic indicators under specific conditions. Specifically, differences are noted in the *boring* and *safe* dimensions under the Gender indicator, and in the *beautiful* and *lively* dimensions under the Major indicator. This

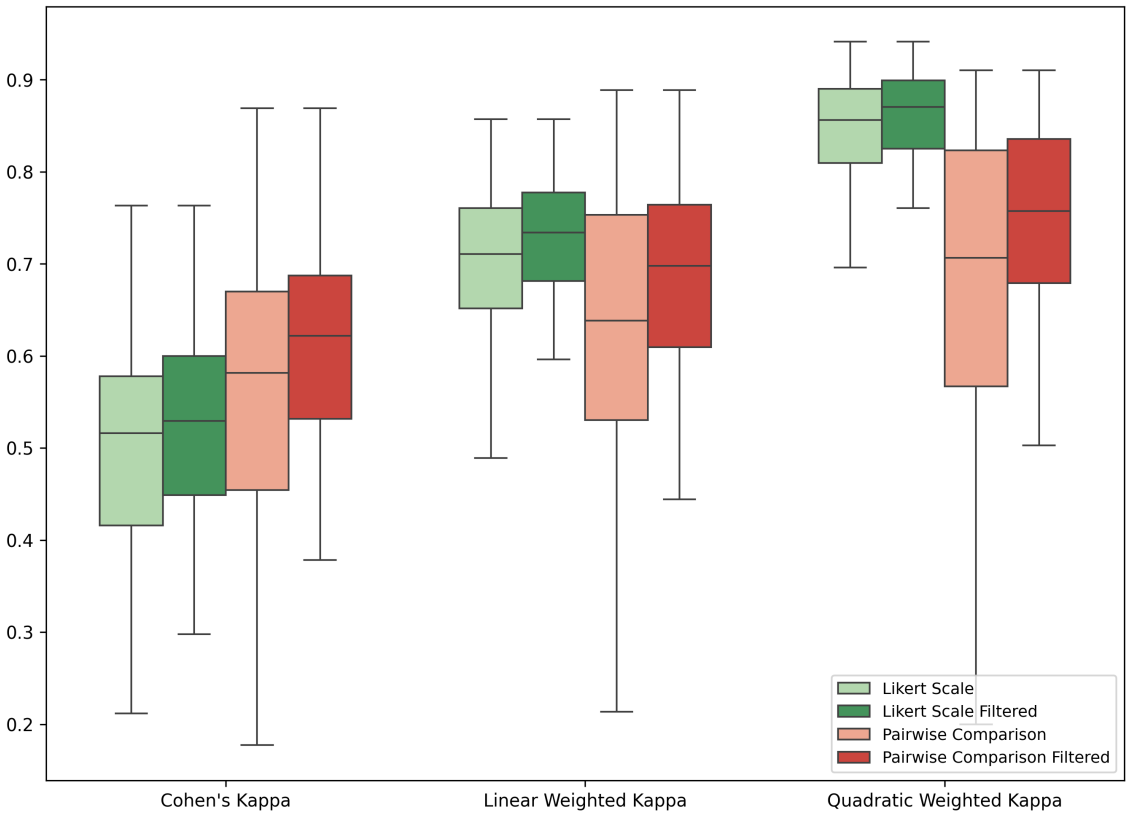

Figure 2: Kappa statistics of intra-rater consistency (the higher the better).

suggests that future surveys should carefully consider the diversity of participants to ensure the representativeness and fairness of the scores. Additionally, this approach will help to more accurately identify the preferences of different groups, thereby avoiding biases caused by a homogeneous demographic background.

Next, Figure 4 compares the correlations of the overall scores (the Mean Score, Q Score, and Trueskill Score) obtained from the two scoring methods. The different scoring methods show strong correlations ($R > 0.8$), indicating robust trends in overall scores across the three scoring calculation methods. In comparing the Q Score and the Trueskill Score, considering the same raw data achieves high correlations ($R > 0.9$) under all parameter settings. In the comparison between the Mean Score and the other two scores, the correlation for Perspective Image is lower, possibly due to amplified cognitive differences in rating and ranking scoring within limited perspectives and scene information.

The t-test results of the survey method duration indicate significant statistical differences in duration between the two methods. Figure 5 shows that Pairwise Comparison takes longer than the Likert Scale, likely because raters need to observe two images instead of one for each score.

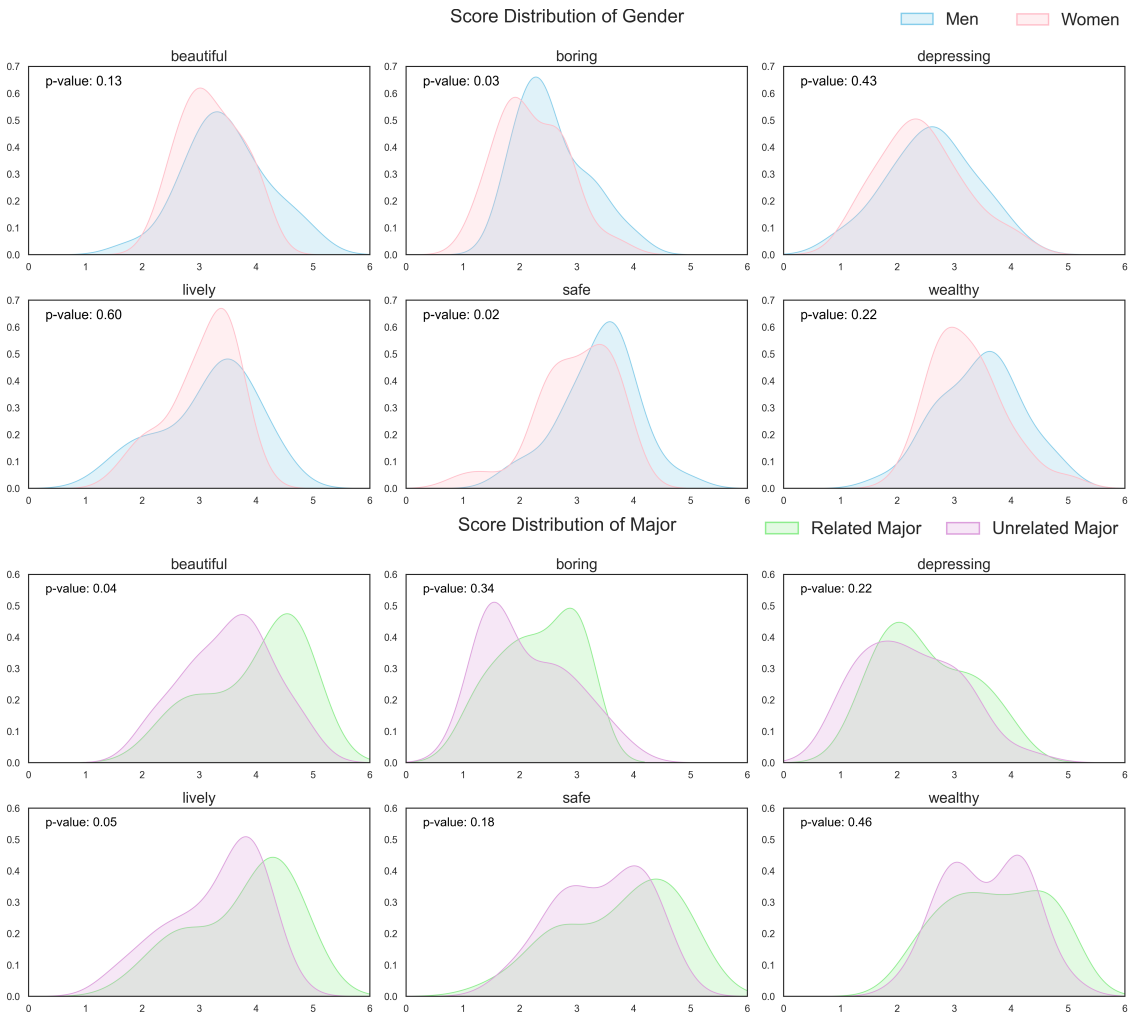

Figure 3: Comparison of overall score distribution across different demographic backgrounds.

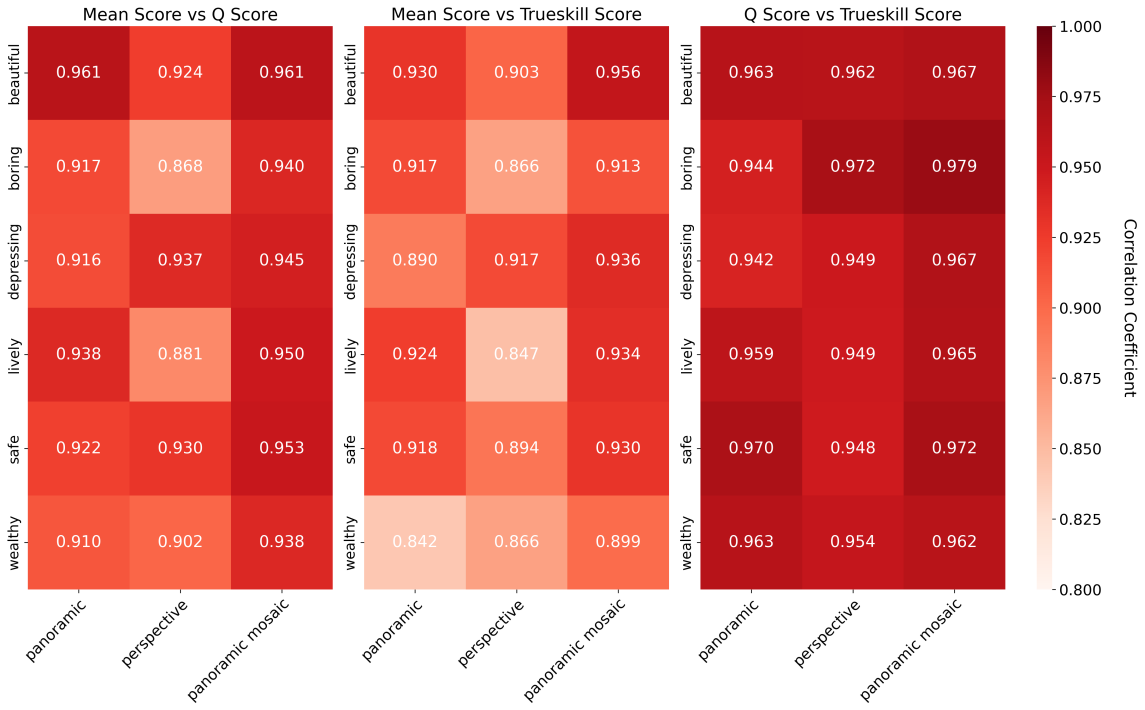

Figure 4: Correlation comparisons of three overall scores (Mean Score, Q Score, and Trueskill Score) from two scoring methods.

Although Pairwise Comparison is typically thought to simplify the decision-making process, this finding suggests that image quantity may have a more substantial impact on total survey duration.

In the comparison of image types, Figure 6 shows that the overall score correlations were generally high (R mostly > 0.7). Panoramic Images and Panoramic Mosaic Images show the highest correlations ($R > 0.8$), higher than those between Perspective Images and the other two. This indicates that Panoramic Images and Panoramic Mosaic Images exhibit the most similar perceptual experiences. Considering that Panoramic Mosaic Images disrupt the original order of image elements, this suggests that the restrictions of perspective and content display have a more significant impact on overall scores than the order of scene elements. Despite this, the consistent overall scores imply that the front view in Perspective Image, even with limited perspective, provides a good representation of scenes.

Specifically, across different parameter settings, the overall scores of the three types of images achieve relatively consistent results across six different indicators and three different scores. The overall consistency of the Trueskill score is lower than that of the other two scores. Given that the raw data for the Trueskill score is the same as that for the Q Score, this discrepancy might be attributed to the higher sensitivity of the Trueskill algorithm to the same data, resulting in greater score fluctuations.

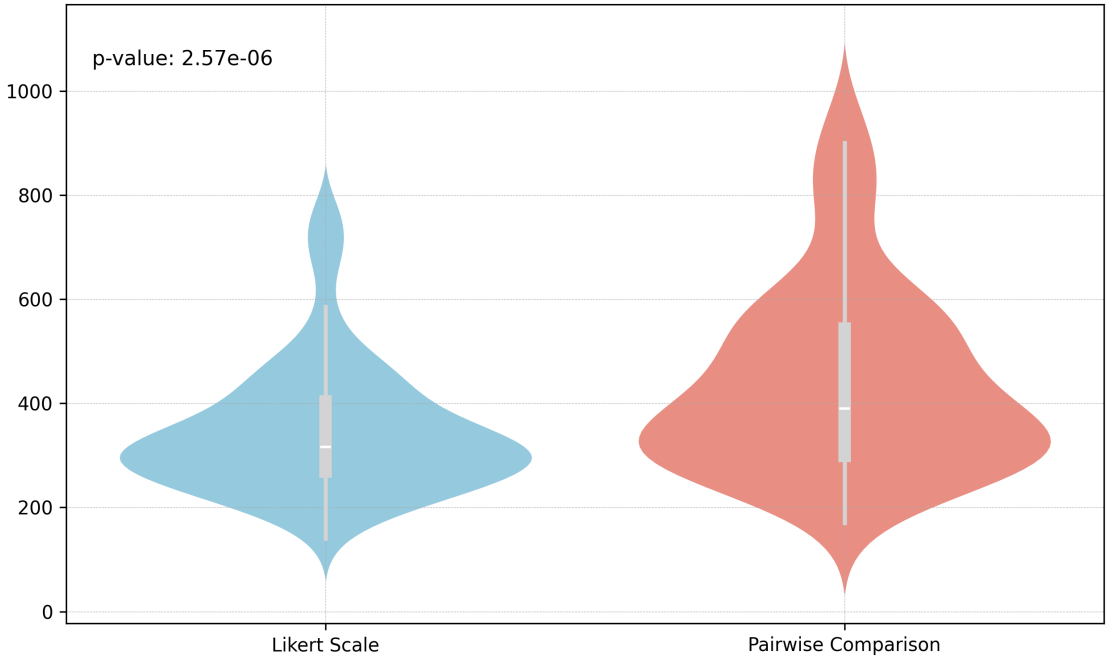

Figure 5: Comparisons of survey duration (seconds) between two scoring methods.

Subsequently, we performed a Mann-Whitney U test, calculating the mean for various image types and indicators across all 180 groups, as illustrated in Figure 7. Overall, the average significance proportion for Pairwise Comparison (P = 0.161, n = 29) is significantly lower than the Likert Scale (P = 0.313, n = 56), indicating that Pairwise Comparison's ranking scoring reduced the differences in image types, focusing more on overall perceptual comparison between images. For image type averages, the pair of 'Panoramic Images vs Panoramic Mosaic Images' has significantly lower significance proportions (Likert Scale = 0.222, Pairwise Comparison = 0.089), consistent with the high correlation of their overall scores, while Perspectives Images show greater variability with the other two types.

In the comparison of different indicators, the *Wealthy* indicator on the Likert scale exhibits a significantly higher proportion compared to other indicators (P = 0.411), whereas in the Pairwise Comparison, it is significantly lower (P = 0.067). This discrepancy may be attributed to the challenge of establishing a unified absolute standard for *Wealthy* among raters, as various factors influence the Likert Scale scores. Conversely, Pairwise Comparison facilitates a consensus on which image appears wealthier. Furthermore, the *Boring* indicator reveals notable differences in significance proportions. On the Likert Scale, the Panoramic Images and Perspective Images categories exhibit the highest proportion (P = 0.500), while the Panoramic Images and Panoramic Mosaic Images categories have the lowest (P = 0.067). This may be due to the limited perspectives

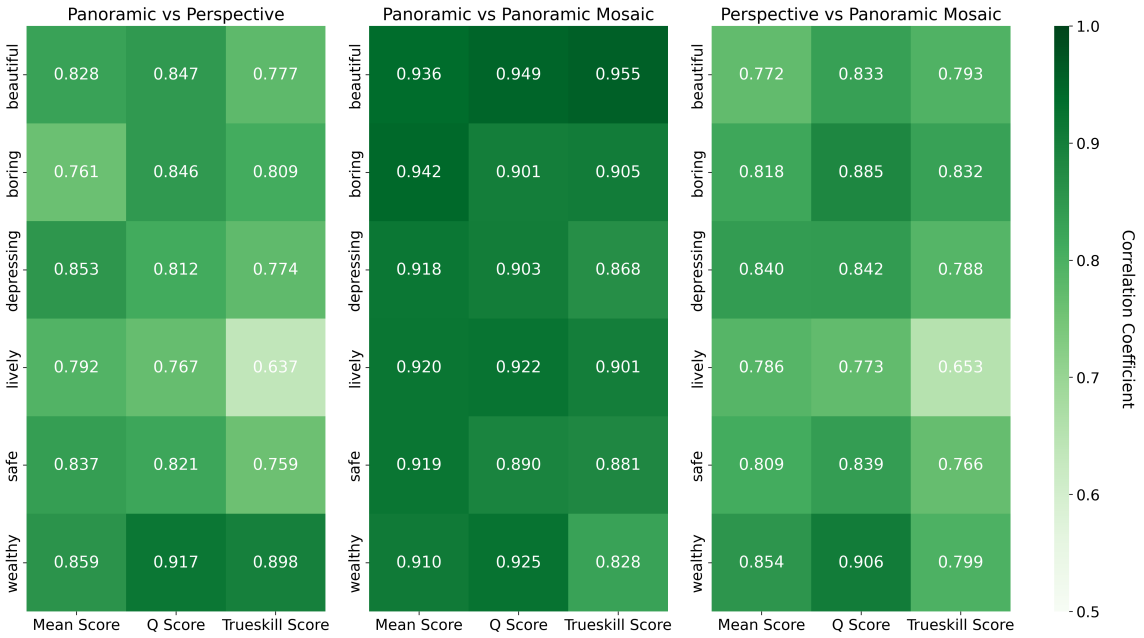

Figure 6: Correlation comparisons of overall scores among three image types.

either emphasizing or diminishing certain features, such as pedestrians or urban furniture, within a smaller view.

Finally, we calculated the kernel density estimation (KDE) and t-test statistics for the standard deviations of raw scores across three image types, as shown in Figure 8. The results reveal significant differences between Perspective Images and the other two types (p-values < 0.01), with Perspective Images exhibiting significantly lower standard deviations. This suggests that Perspective Images yield more stable subjective scores within limited perspectives and elements. As visual perception research increasingly emphasizes not only overall image scores but also the distribution and deviation of scores, the choice of image type and scoring method becomes crucial. Panoramic and Panoramic Mosaic Images reflect the complete element distribution of a scene, while Perspective Images provide more stable and reliable subjective scores.

4.3. Robustness

We adopted a consistent robustness calculation strategy for three different scores (the Mean Score, Q score and Trueskill Score) to facilitate comparison. The number of ratings of Likert Scale data was slightly smaller than that of Pairwise Comparison, with the maximum number set at 30 and 45 times, respectively. We conducted separate calculations for different locations and image types, summarizing the average values through each indicator.

We calculated the rating quantity range corresponding to six indicators at a robustness level of 0.8, setting the minimum value as the conservative value and the maximum value as the stable

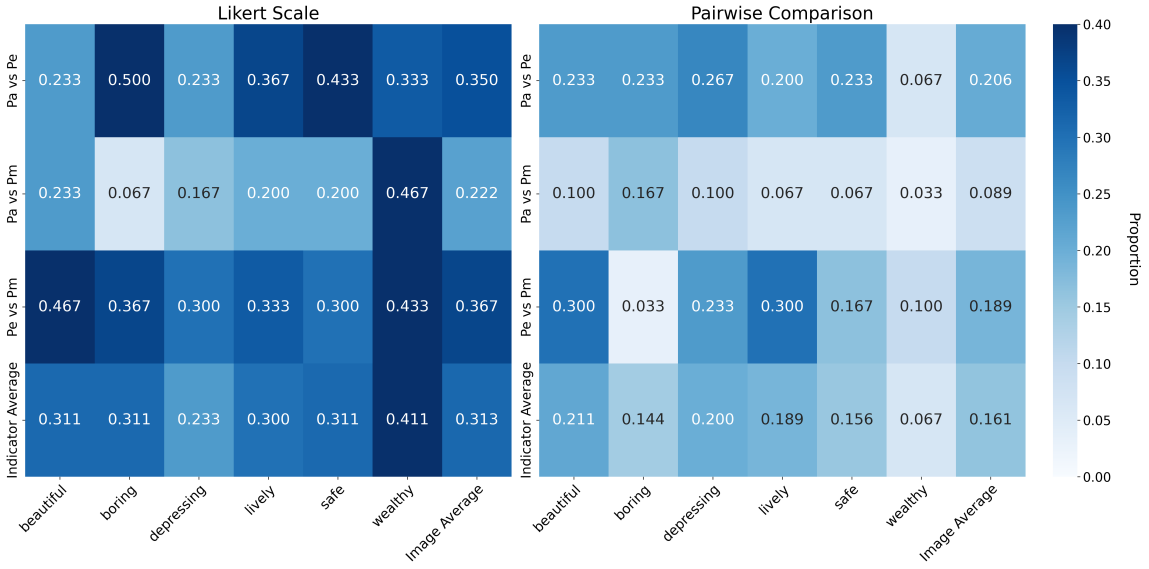

Figure 7: Comparisons of raw scores among three image types. (*Pa* denotes Panoramic Images, *Pe* denotes Perspective Images, and *Pm* denotes Panoramic Mosaic Images.)

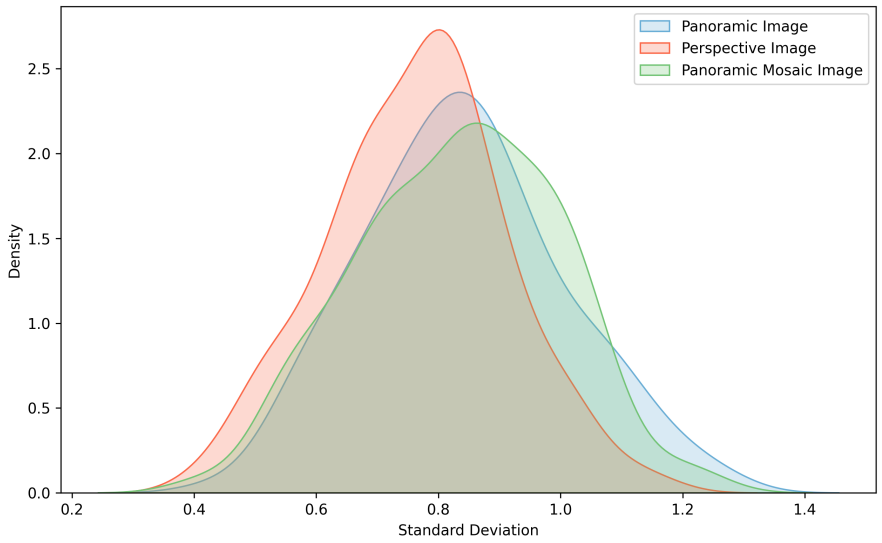

Figure 8: Kernel density estimation of the standard deviation among three image types.

value. As illustrated in Figure 9, the Likert Scale (the Mean Score) demonstrates the fastest convergence speed, with a conservative value of 12 times and a stable value of 19 times. Pairwise Comparison (the Q Score) shows slower convergence, with conservative and stable values of 22

and 29 times, respectively. Pairwise Comparison (the Trueskill Score) has the slowest convergence, with conservative and stable values of 29 and 37 times, respectively. Overall, to achieve the same standard of robustness, Pairwise Comparison requires significantly more scoring times than the Likert Scale. Although each scoring in Pairwise Comparison produces scores for two images, the previously confirmed longer scoring duration means it cannot be simply halved.

It is noteworthy that the stability of the Likert Scale scores is lower than that of Pairwise Comparison. Within individual indicators, the variability of robustness is significantly larger for Likert Scale scores compared to its counterpart. Additionally, there are notable discrepancies in the range of scores across different metrics. This may be due to the biases in human consensus introduced by the rating scoring method and the relatively simple scoring calculation method, which amplifies the cognitive differences among raters. In contrast, Pairwise Comparison's ranking scoring helps raters form more stable judgments, resulting in smoother and more similar scores. Interestingly, the *Beautiful* indicator showed the fastest convergence across all scoring methods, indicating strong human consensus on this attribute. Conversely, for indicators like Wealthy, which are challenging to establish absolute standards, the Likert Scale had slower convergence, whereas Pairwise Comparison did not exhibit this bias.

To verify the credibility of our robustness analysis, we conducted a comparable analysis using another open dataset, SPECS (Street Perception Evaluation Considering Socioeconomics), which was developed by our research group independently of this project. This dataset spans a diverse demographic including 1,000 residents across five continents, and captures human perceptions of streetscapes based on ten indicators, six of which are the same in this study (see https://github.com/matqr/specs). We selected over 50 images, each with more than 30 ratings on the same six indicators. Since this dataset was obtained via Pairwise Comparison, only the Q score and Trueskill score were computed. For consistency with our previous analysis, we calculated the first 15 data points (largest non-overlapping group for 30 ratings) and fitted the subsequent data points as shown in Figure 10. The validation results closely match our earlier findings. Specifically, the Q score showed conservative and stable thresholds at 20 and 36 times, respectively, with a larger range, possibly due to variability in data quality or curve-fitting errors. The Trueskill score demonstrated nearly identical thresholds at 27 and 34 times. Thus, accounting for some errors, our proposed conservative and stable number ranges achieve a robustness level of approximately 0.8 across another diverse demographic dataset. This finding not only demonstrates the broader applicability of our results across different population groups but also provides additional empirical evidence supporting the consistency of perception patterns between student populations and the general public.

Based on this finding, we recognize that while many image-based visual perception studies meet the required conservative number of scoring votes (Lu and Chen, 2024; Ogawa et al., 2024; Chen and Biljecki, 2023; Qiu et al., 2023), a significant portion still suffers from insufficient raters or scoring votes (Rui and Cai, 2025; Ito and Biljecki, 2021; Kang et al., 2023; Luo et al., 2022b; Kruse et al., 2021). Additionally, some studies fail to report the relevant sample size, which hinders the ability to assess the reliability of the evaluation. This issue can significantly impact representativeness, especially considering that many studies use survey data as training datasets to obtain

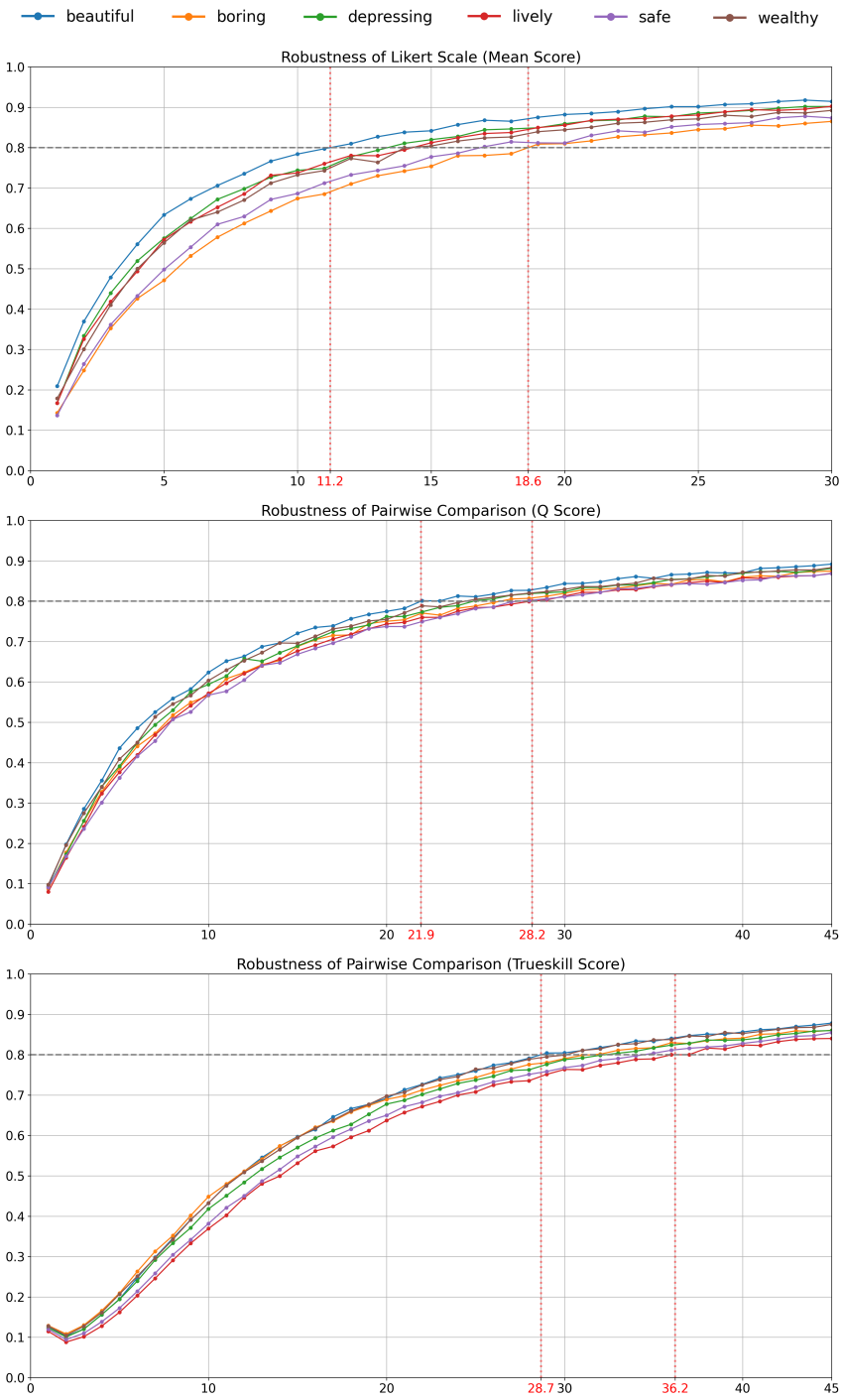

Figure 9: Robustness of rating number under three scores.

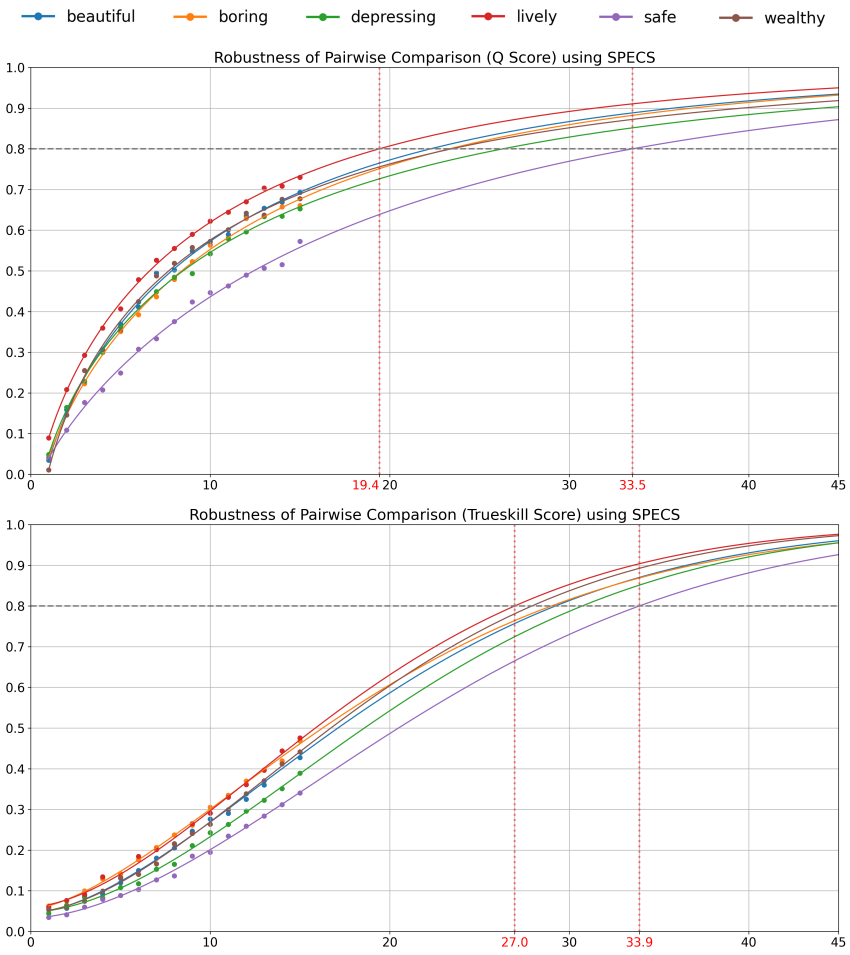

Figure 10: Robustness of rating number using SPECS.

subjective indicator score distributions widely. If the quality of upstream data is compromised, it may lead to greater error propagation in downstream tasks. Therefore, we recommend researchers meet the conservative value recommendations for different methods as much as possible and enhance their statistical data validation. Furthermore, considering the widespread use of the Likert Scale in current visual perception surveys, we encourage a shift towards Pairwise Comparison whenever there are sufficient resources to secure rater numbers and time, producing more stable scores and smaller discrepancies between metrics.

5. Discussion and guidelines

5.1. Main insights and recommendations

Based on the aforementioned insights, we propose future survey design recommendations in four key areas: survey structure, raters selection, scoring method selection, and image selection. Additionally, we highlight the importance of applying these recommendations in machine learning tasks, a critical downstream application of visual perception surveys.

Regarding survey structure, a between-subject design, such as the non-overlapping group strategy, has proven effective in achieving robust survey results and facilitating the filtering of individual data noise to attain a substantial agreement level (Weighted Kappa > 0.7). In resource-limited scenarios, where only a single survey round can be conducted, applying these strategies can enhance data stability and reliability.

In terms of raters selection, a potential statistical impact was found with the limited number of raters and subjective indicators. It remains uncertain whether these biases would be amplified in larger or unbalanced groups. Therefore, to obtain more stable and statistically representative results, it is recommended to select individuals from diverse backgrounds with balanced quantities whenever possible. Another crucial parameter is sampling. Researchers are encouraged to estimate the sample size based on the number of assessed images before conducting the survey. Specifically, to reach a reliable and reproducible survey, the number of ratings per image should not be less than the conservative estimates set: for each indicator, 12 times for the Likert Scale (the Mean Score), and 22 times and 29 times for the Q Score and the Trueskill Score in Pairwise Comparison, respectively. Notably, a recent study indicates that despite strong alignment between AI models and human evaluations, significant geographic variations persist. Consequently, human raters remain essential for such tasks in the foreseeable future (Malekzadeh et al., 2025).

Concerning scoring method selection, both the Likert Scale and Pairwise Comparison exhibit comparable trends in overall scores (R mostly > 0.7), but each has distinct characteristics suitable for different contexts. As a rating scoring method, the Likert Scale, despite potential issues with standardization and data fluctuation in specific scoring options, offers shorter scoring durations and fewer minimum scoring ratings required. This makes it an efficient scoring method for small to medium-sized surveys with limited resources and time. Its parametric nature also allows for various parametric mathematical and statistical analyses of the raw scoring data, providing higher explanatory power. In contrast, the Pairwise Comparison method, a ranking scoring approach, demonstrates better inter-rater consistency (Kappa = 0.589). It exhibits a smaller significant proportion of raw scores across various image types (P = 0.161) and similar convergence rates among different subjective indicators in robustness calculations, indicating a more stable scoring method. However, it generally requires more ratings and time, making it suitable for surveys with ample resources and extended timelines. Considering that some large-scale visual perception surveys currently use the Likert Scale (Fitch et al., 2022; Suppakittpaisarn et al., 2020), it is advisable to consider adopting Pairwise Comparison in the future, when resources allow, to obtain more stable scores.

With respect to image type selection, three common image types — Panoramic Images, Perspective Images (front view), and Panoramic Mosaic Images — exhibit different trends. If the goal is to obtain overall scores of images through surveys, a strong correlation ($R > 0.8$) indicates that different image contents exhibit highly similar scoring trends, allowing for a certain degree of interchangeability. Considering that current image types are primarily manually taken, non-geotagged photos, and commercial SVI, these data sources often suffer from long collection times, limited coverage and potential future restrictions. Shifting to crowdsourced data sources composed mainly of Perspective Images (such as Mapillary or KartaView) might be a better option. These images are not confined by coverage and drivable roads and can be collected by different users in alleys, parks, and even water bodies, greatly expanding the image boundaries (Chen and Biljecki, 2023; Luo et al., 2022b). Coupled with Biljecki et al. (2023)'s findings that Perspective Images can serve as good representatives for analyzing objective indicators (e.g., assessing the amount of greenery in a street), this discovery reaffirms the significant value of crowdsourcing imagery in urban research. Leveraging survey data or open datasets to train machine learning models now makes it possible to conduct large-scale assessments of urban visual perception in areas not covered by commercial SVIs. If considering the raw scoring records of different images, an increasing number of subjective studies based on SVI are not limited to the overall score of images but explore the distribution, preferences, and deviation of scoring data of different populations during the process (Cui et al., 2023; Kang et al., 2023; Meir and Oron-Gilad, 2020; Hidayati et al., 2020). There is a considerable discrepancy among the three types of images, which is further amplified in the Likert Scale. Perspective images have been shown to exhibit greater scoring stability, consistent with the findings of Beaucamp et al., possibly due to their limited scene elements, whereas the other two types of images contain more comprehensive scene information. In exploratory data tasks, the choice of image type requires further consideration and selection.

As an important downstream task in urban visual perception surveys, machine learning techniques are often deployed to map visual perception scores at the city level and analyze the specific impact of explanatory variables on human perception (Zhao et al., 2023; Zhang et al., 2018; Chen and Biljecki, 2023; Kruse et al., 2021; Yao et al., 2019; Ye et al., 2019). From this perspective, conducting effective surveys becomes particularly crucial. First, it ensures data quality. For instance, by meeting the minimum rating threshold, we can ensure robustness across different demographic groups. The use of robust survey structures such as between-subject design, and more stable scoring methods such as Pairwise Comparison, helps reduce noise and variability in the data, leading to more consistent and accurate input for machine learning algorithms. These practices contribute to more representative scores, minimize error propagation, and improve model generalizability in subsequent analysis. Second, effective surveys facilitate model benchmarking. Since visual perception is applied in various contexts, such as housing price assessment (Kang et al., 2021), identifying inconspicuous places (Zhang et al., 2020), cycling volume evaluation (Gao and Fang, 2025), visual environmental quality assessment (Wang et al., 2024) and so on, the previous survey data collection processes have varied significantly. By standardizing these practices, future models can be compared more fairly, thus contributing to more reliable benchmarking.

Table 2: Proposal of survey reporting parameters and examples.

Section	Reporting Parameters	Example
Survey Structure	Design, Allocation, Consistency evaluation	We established between-subject random non-overlapping subsets to assess rater data consistency across different groups and to filter noise through confidence intervals.
Raters	Demographics, Sample size, Number of ratings per image, Recruitment process	We recruited 500 undergraduate students from various majors at our university, all of whom are Singaporean, with an average age of 20 years and a gender ratio of 1:1. Each participant completed 100 ratings, with each image receiving an average of 50 ratings per indicator, fulfilling the stable rating requirements.
Scoring	Scoring methodology, Rationale for method selection, Evaluative indicators	We selected Pairwise Comparison as the ranking scoring method due to its proven stability over rating methods while meeting time and rating needs. Participants in the graphical interface were asked to choose which of the two images was safer, more boring, or equally rated.
Image	Image type, Image source, Rationale for image choice, Dataset composition	We chose Perspective Images for their reliable subjective rating stability and easy availability from open data sources. Using Mapillary as the data source, we used 500 images from Singapore, for subjective ratings in the survey.

5.2. Reporting protocol

Several key parameters are frequently omitted or underreported in existing studies. For example, although the number of raters is commonly provided, the number of ratings per image is rarely stated. Therefore, in addition to design recommendations, we propose establishing a reporting protocol that specifies and communicates the key parameters and contributes to the understanding of survey design. In Table 2, we summarize the four key sections of visual survey design along with their corresponding parameters and concise examples to illustrate the format for reporting and elaborating the motivation for these elements. With this protocol, we hope to encourage researchers to plainly report these details, thereby promoting more transparent, credible, and reproducible experiments and ultimately setting a standard.

5.3. Limitations and future work

One limitation of this study lies in the selection of survey parameters. To balance survey resources and time, we chose the two most common scoring methods from web-based surveys and three types of SVI data. However, the field offers additional avenues for valuable insights. Alternative scoring platforms and techniques such as eye-tracking (Li et al., 2020; Yang et al., 2024), EEG (Mavros et al., 2022), and wearables (Luo et al., 2024) provide a diverse way to capture human perception. Even within our chosen scoring methods, variations exist. For instance, an enhanced TrueSkill algorithm that includes temporal and spatial effects has demonstrated improved reliability in a certain task, potentially reducing the samples required compared to our findings (Qiu

et al., 2025). Additionally, other visual data formats beyond street road perspectives — such as waterscapes (Luo et al., 2022a), pedestrian walkways (Chen et al., 2024), and building facades (Liang et al., 2024) — provide further insights. Future studies could explore these specific parameters across different tasks while prioritising sampling, parameter justification, and reporting, as proposed in this study, to enhance the reliability and reproducibility of research findings in the field. Furthermore, it remains crucial to investigate whether various visual data types and scoring methods accurately reflect human perception in real environments. This will help assess the representativeness of different methodologies in practical settings.

In terms of sample selection, our survey also has certain limitations. Due to resource constraints, we selected a relatively homogeneous group within a university. While we believe the analysis based on this sample is valid, given that similar practices have never been conducted in this field, thus it serves as a blueprint and underscores the importance of considering sample selection in ensuring survey reliability. This approach also aligns with many current studies in the field that focus on specific groups (Qiu et al., 2023; Ye et al., 2019; Kawshalya et al., 2022; Tang and Long, 2019; Li et al., 2022; Gong et al., 2023; Ma et al., 2023). Although we validated key factors such as the number of ratings required to reach a robustness level on a diverse dataset, caution is needed when extending these conclusions to more heterogeneous groups, such as specific genders (e.g., women) or age groups (e.g., the elderly). In future studies, expanding the diversity of the sample and conducting comprehensive analyses on the impact of different demographic factors, such as nationality, income, and personality, on perception research could be a promising direction.

6. Conclusion

Riding the wave of increasing visual data availability, digital connectivity, and computational advances, perception studies in urban planning and design have scaled and revealed insights previously out of reach. Now we can better understand human perception of urban environments and identify potential influencing factors such as demographic characteristics (Kazemi et al., 2023; Gong et al., 2023), image physical visual features (Kawshalya et al., 2022; Li et al., 2022) and urban functional layout (Zeng et al., 2024). While the studies are now conducted routinely and increasing in volume, it seems that they have not fully transcended their formative period — there is a lack of standards and understanding of the availability, reliability, and validity of approaches.

Using imagery to measure objective information with techniques such as semantic segmentation, e.g., quantifying greenery, is consistent, not leaving much room for ambiguity. In contrast, image-based visual perception studies face difficulties in forming unified indicators due to the diverse goals across different domains, which is compounded by the underlying human subjectiveness. Therefore, researchers typically design their human surveys and recruit raters to collect perception data based on factors such as budget, time, survey platform, and the availability of visual data. This melange leads to varying survey parameters and a lack of references for selecting appropriate parameters.

For the first time, our study offers a comprehensive and domain-agnostic framework, supporting the research community in establishing a consensus on survey design and clarifying many

uncertainties that have long puzzled the field. These parameters cover aspects commonly involved in visual perception surveys, including structure, raters, scoring methods, and image types, offering insights and recommendations tailored to different applications. Our study has delivered several takeaways, among which the three key recommendations are as follows.

Raters: Many previous studies overlooked reporting the number of ratings per image, which is of utmost importance, while some failed to meet the requirements for robustness (e.g., Mean Score = 12, Q Score = 22, Trueskill Score = 29). This parameter significantly drives the quality of studies, with sparse data being deficient in representativeness and introducing bias, possibly propagating unsound conclusions. Therefore, we encourage researchers to meet the conservative rating values per image for each indicator we propose to ensure robust ratings and to report this parameter to enhance credibility and reproducibility.

Scoring Methods: The two typical scoring methods — the rating method and the ranking method — exhibit similar trends in overall scores. However, each method has different applications. For example, the Likert Scale is suitable for smaller surveys constrained by resources (e.g., participants, funding, time) and facilitates parameterized statistical analysis, while Pairwise Comparison is suitable for abundant surveys and provides more robust scores. We encourage researchers to determine specific scoring schemes based on their objectives and to report detailed survey designs. Considering that current surveys heavily rely on the Likert Scale, we also encourage researchers to consider Pairwise Comparison when possible for stability.

Image Types: Different types of images exhibit comparable trends in representing overall scene perception, indicating that crowdsourced SVI, primarily composed of non-panoramas (Biljecki et al., 2023), shows great potential for use. Considering that most current imagery is based on manually obtained non-geo-tagged photos and commercial SVI from companies such as Google and Baidu, manual image acquisition is time-consuming, and it is uncertain whether commercial street view data will remain easily accessible to researchers (Helbich et al., 2024). Therefore, in the context of the rapid development in crowdsourcing, we encourage researchers to consider using crowdsourced imagery (e.g., see the recent open dataset by Hou et al. (2024)) and expand the boundaries of urban science.

Author contributions

Youlong Gu — Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft; Matias Quintana — Conceptualization, Methodology, Project administration, Writing – review and editing; Xiucheng Liang — Conceptualization, Methodology, Writing – review and editing; Koichi Ito — Conceptualization, Methodology, Writing – review and editing; Winston Yap — Conceptualization, Methodology, Writing – review and editing; Filip Biljecki — Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review and editing.

Acknowledgements

We gratefully acknowledge the participants of the experiments. The research was approved by the Ethical Review Committee at our department (reference code: 2024/DOA/001). We thank our colleagues at the NUS Urban Analytics Lab for their insightful discussions, especially Yujun Hou for the valuable suggestions. This research is part of the projects (i) Large-scale 3D Geospatial Data for Urban Analytics, which is supported by the National University of Singapore under the Start Up Grant R-295-000-171-133 and (ii) Multi-scale Digital Twins for the Urban Environment: From Heartbeats to Cities, which is supported by the Singapore Ministry of Education Academic Research Fund Tier 1. The research was partially conducted at the Future Cities Lab Global at the Singapore-ETH Centre, which was established collaboratively between ETH Zürich and the National Research Foundation Singapore (NRF) under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The research was supported by the Singapore International Graduate Award Scholarship.

References

Aithal, A., Aithal, P.S., 2020. Development and Validation of Survey Questionnaire & Experimental Data – A Systematical Review-based Statistical Approach doi:10.5281/ZENODO.4179499.

Ak, A., Abid, M., Silva, M.P.D., Callet, P.L., 2021. On Spammer Detection In Crowdsourcing Pairwise Comparison Tasks: Case Study On Two Multimedia Qoe Assessment Scenarios, in: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6. doi:10.1109/ICMEW53276.2021.9455992.

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google Street View: Capturing the World at Street Level. IEEE Computer 43, 32–38. doi:10.1109/MC.2010.170.

Artino, A.R.J., Durning, S.J., Sklar, D.P., 2018. Guidelines for Reporting Survey-Based Research Submitted to Academic Medicine. Academic Medicine 93, 337. doi:10.1097/ACM.0000000000002094.

Beaucamp, B., , Thomas, L., , Vincent, T., , Servières, M., . Beyond the frame: evaluating panoramic vs. perspective images for assessing place perception. International Journal of Geographical Information Science 0, 1–33. doi:10.1080/13658816.2025.2483857.

Beaucamp, B., Leduc, T., Tourre, V., Servières, M., 2022. THE WHOLE IS OTHER THAN THE SUM OF ITS PARTS: SENSIBILITY ANALYSIS OF 360° URBAN IMAGE SPLITTING. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-4-2022, 33–40. doi:10.5194/isprs-annals-V-4-2022-33-2022.

Bennett, C., Khangura, S., Brehaut, J.C., Graham, I.D., Moher, D., Potter, B.K., Grimshaw, J.M., 2011. Reporting Guidelines for Survey Research: An Analysis of Published Guidance and Reporting Practices. PLOS Medicine 8, e1001069. doi:10.1371/journal.pmed.1001069.

Bethlehem, J., 2009. Applied Survey Methods: A Statistical Perspective.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. Landscape and Urban Planning 215, 104217. doi:10.1016/j.landurbplan.2021.104217.

Biljecki, F., Zhao, T., Liang, X., Hou, Y., 2023. Sensitivity of measuring the urban form and greenery using street-level imagery: A comparative study of approaches and visual perspectives. International Journal of Applied Earth Observation and Geoinformation 122, 103385. doi:10.1016/j.jag.2023.103385.

Brush, R., Chenoweth, R.E., Barman, T., 2000. Group differences in the enjoyability of driving through rural landscapes. Landscape and Urban Planning 47, 39–45. doi:10.1016/S0169-2046(99)00073-0.

Cameron, J.J., Stinson, D.A., 2019. Gender (mis)measurement: Guidelines for respecting gender diversity in psychological research. Social and Personality Psychology Compass 13, e12506. doi:10.1111/spc3.12506.

Cao, Y., Yang, P., Xu, M., Li, M., Li, Y., Guo, R., 2025. A novel method of urban landscape perception based on biological vision process. Landscape and Urban Planning 254, 105246. doi:10.1016/j.landurbplan.2024.105246.

Chen, C., Li, H., Luo, W., Xie, J., Yao, J., Wu, L., Xia, Y., 2022. Predicting the effect of street environment on residents' mood states in large urban areas using machine learning and street view images. Science of The Total Environment 816, 151605. doi:10.1016/j.scitotenv.2021.151605.

Chen, S., Biljecki, F., 2023. Automatic assessment of public open spaces using street view imagery. Cities 137, 104329. doi:10.1016/j.cities.2023.104329.

Chen, Y., Huang, X., White, M., 2024. A study on street walkability for older adults with different mobility abilities combining street view image recognition and deep learning - The case of Chengxianjie Community in Nanjing (China). Computers, Environment and Urban Systems 112, 102151. doi:10.1016/j.compenvurbsys.2024.102151.

Clay, G.R., Smidt, R.K., 2004. Assessing the validity and reliability of descriptor variables used in scenic highway analysis. Landscape and Urban Planning 66, 239–255. doi:10.1016/S0169-2046(03)00114-2.

Cornsweet, T., 2012. Visual Perception.

Cui, Q., Zhang, Y., Yang, G., Huang, Y., Chen, Y., 2023. Analysing gender differences in the perceived safety from street view imagery. International Journal of Applied Earth Observation and Geoinformation 124, 103537. doi:10.1016/j.jag.2023.103537.

Dell, R.B., Holleran, S., Ramakrishnan, R., 2002. Sample Size Determination. ILAR Journal 43, 207–213. doi:10.1093/ilar.43.4.207.

Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A., 2016. Deep Learning the City : Quantifying Urban Perception At A Global Scale.

Eng, J., 2003. Sample Size Estimation: How Many Individuals Should Be Studied?1. Radiology .

Fink, A., 2003. The Survey Handbook.

Fitch, D.T., Carlen, J., Handy, S.L., 2022. What makes bicyclists comfortable? Insights from a visual preference survey of casual and prospective bicyclists. Transportation Research Part A: Policy and Practice 155, 434–449. doi:10.1016/j.tra.2021.11.008.

Gadermann, A.M., Guhn, M., Zumbo, B.D., 2012. Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. Practical Assessment, Research, and Evaluation 17. doi:10.7275/n560-j767.

Gao, M., Fang, C., 2025. Pedaling through the cityscape: Unveiling the association of urban environment and cycling volume through street view imagery analysis. Cities 156, 105573. doi:10.1016/j.cities.2024.105573.

Gao, T., Liang, H., Chen, Y., Qiu, L., 2019. Comparisons of Landscape Preferences through Three Different Perceptual Approaches. International Journal of Environmental Research and Public Health 16, 4754. doi:10.3390/ijerph16234754.

Gibson, J.J., 2014. The Ecological Approach to Visual Perception: Classic Edition. New York. doi:10.4324/9781315740218.

Gong, W., Huang, X., White, M., Langenheim, N., 2023. Walkability Perceptions and Gender Differences in Urban Fringe New Towns: A Case Study of Shanghai. Land 12, 1339. doi:10.3390/land12071339.

Groves, R.M., Jr, F.J.F., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R., 2011. Survey Methodology.

Hallgren, K.A., 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in quantitative methods for psychology 8, 23–34.

Harpe, S.E., 2015. How to analyze Likert and other rating scale data. Currents in Pharmacy Teaching and Learning 7, 836–850. doi:10.1016/j.cptl.2015.08.001.

Helbich, M., Danish, M., Labib, S.M., Ricker, B., 2024. To use or not to use proprietary street view images in (health and place) research? That is the question. Health & Place 87, 103244. doi:10.1016/j.healthplace.2024.103244.

Herbrich, R., Minka, T., Graepel, T., 2006. TrueSkill™: a Bayesian skill rating system, in: Proceedings of the 19th International Conference on Neural Information Processing Systems, Cambridge, MA, USA. pp. 569–576.

Hidayati, I., Tan, W., Yamu, C., 2020. How gender differences and perceptions of safety shape urban mobility in Southeast Asia. Transportation Research Part F: Traffic Psychology and Behaviour 73, 155–173. doi:10.1016/j.trf.2020.06.014.

Hou, Y., Biljecki, F., 2022. A comprehensive framework for evaluating the quality of street view imagery. International Journal of Applied Earth Observation and Geoinformation 115, 103094. doi:10.1016/j.jag.2022.103094.

Hou, Y., Quintana, M., Khomiakov, M., Yap, W., Ouyang, J., Ito, K., Wang, Z., Zhao, T., Biljecki, F., 2024. Global Streetscapes — A comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. ISPRS Journal of Photogrammetry and Remote Sensing 215, 216–238. doi:10.1016/j.isprsjprs.2024.06.023.

Huang, G., Yu, Y., Lyu, M., Sun, D., Zeng, Q., Bart, D., 2023. Using google street view panoramas to investigate the influence of urban coastal street environment on visual walkability. Environmental Research Communications 5, 065017. doi:10.1088/2515-7620/acdecf.

Hung, S.H., Chang, C.Y., 2022. How do humans value urban nature? Developing the perceived biophilic design scale (PBDs) for preference and emotion. Urban Forestry & Urban Greening 76, 127730. doi:10.1016/j.ufug.2022.127730.

Ito, K., Biljecki, F., 2021. Assessing bikeability with street view imagery and computer vision. Transportation Research Part C: Emerging Technologies 132, 103371. doi:10.1016/j.trc.2021.103371.

Ito, K., Kang, Y., Zhang, Y., Zhang, F., Biljecki, F., 2024. Understanding urban perception with visual data: A systematic review. Cities 152, 105169. doi:10.1016/j.cities.2024.105169.

Jebb, A.T., Ng, V., Tay, L., 2021. A Review of Key Likert Scale Development Advances: 1995–2019. Frontiers in Psychology 12. doi:10.3389/fpsyg.2021.637547.

Kang, L., Xiong, Y., Mannering, F.L., 2013. Statistical analysis of pedestrian perceptions of sidewalk level of service in the presence of bicycles. Transportation Research Part A: Policy and Practice 53, 10–21. doi:10.1016/j.tra.2013.05.002.

Kang, Y., Abraham, J., Ceccato, V., Duarte, F., Gao, S., Ljungqvist, L., Zhang, F., Näsman, P., Ratti, C., 2023. Assessing differences in safety perceptions using GeoAI and survey across

neighbourhoods in Stockholm, Sweden. Landscape and Urban Planning 236, 104768. doi:10.1016/j.landurbplan.2023.104768.

Kang, Y., Zhang, F., Gao, S., Peng, W., Ratti, C., 2021. Human settlement value assessment from a place perspective: Considering human dynamics and perceptions in house price modeling. Cities 118, 103333. doi:10.1016/j.cities.2021.103333.

Kasunic, M., 2005. Designing an effective survey.

Kawshalya, L.W.G., Weerasinghe, U.G.D., Chandrasekara, D.P., 2022. The impact of visual complexity on perceived safety and comfort of the users: A study on urban streetscape of Sri Lanka. PLOS ONE 17, e0272074. doi:10.1371/journal.pone.0272074.

Kazemi, F., Hosseinpour, N., Ebrahimian, M., 2023. People's preferences and perceptions toward low-input versus conventional park design approaches using 3D images and interview-based questionnaires. Urban Forestry & Urban Greening 86, 128040. doi:10.1016/j.ufug.2023.128040.

Kelly, C.M., Wilson, J.S., Baker, E.A., Miller, D.K., Schootman, M., 2013. Using Google Street View to Audit the Built Environment: Inter-rater Reliability Results. Annals of Behavioral Medicine 45, S108–S112. doi:10.1007/s12160-012-9419-9.

Krosnick, J.A., 1999. SURVEY RESEARCH. Annual Review of Psychology 50, 537–567. doi:10.1146/annurev.psych.50.1.537.

Kruse, J., Kang, Y., Liu, Y.N., Zhang, F., Gao, S., 2021. Places for play: Understanding human perception of playability in cities using street view images and deep learning. Computers, Environment and Urban Systems 90, 101693. doi:10.1016/j.compenvurbsys.2021.101693.

Lenth, R.V., 2001. Some Practical Guidelines for Effective Sample Size Determination. The American Statistician 55, 187–193. doi:10.1198/000313001317098149.

Li, J., Zhang, Z., Jing, F., Gao, J., Ma, J., Shao, G., Noel, S., 2020. An evaluation of urban green space in Shanghai, China, using eye tracking. Urban Forestry & Urban Greening 56, 126903. doi:10.1016/j.ufug.2020.126903.

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, 104140. doi:10.1016/j.scs.2022.104140.

Liang, X., Chang, J.H., Gao, S., Zhao, T., Biljecki, F., 2024. Evaluating human perception of building exteriors using street view imagery. Building and Environment 263, 111875. doi:10.1016/j.buildenv.2024.111875.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 22 140, 55–55.

Lis, A., Zalewska, K., Pardela, L., Adamczak, E., Cenarska, A., Bławicka, K., Brzegowa, B., Matiiuk, A., 2022. How the amount of greenery in city parks impacts visitor preferences in the context of naturalness, legibility and perceived danger. Landscape and Urban Planning 228, 104556. doi:10.1016/j.landurbplan.2022.104556.

Liu, Y., Chen, T., Chung, H., Jang, K., Xu, P., 2025. Is there an emotional dimension to road safety? A spatial analysis for traffic crashes considering streetscape perception and built environment. Analytic Methods in Accident Research , 100374doi:10.1016/j.amar.2025.100374.

Lu, Y., Chen, H.M., 2024. Using google street view to reveal environmental justice: Assessing public perceived walkability in macroscale city. Landscape and Urban Planning 244, 104995. doi:10.1016/j.landurbplan.2023.104995.

Luo, J., Zhao, T., Cao, L., Biljecki, F., 2022a. Semantic Riverscapes: Perception and evaluation of linear landscapes from oblique imagery using computer vision. Landscape and Urban Planning 228, 104569. doi:10.1016/j.landurbplan.2022.104569.

Luo, J., Zhao, T., Cao, L., Biljecki, F., 2022b. Water View Imagery: Perception and evaluation of urban waterscapes worldwide. Ecological Indicators 145, 109615. doi:10.1016/j.ecolind.2022.109615.

Luo, L., Jiang, B., 2022. From oppressiveness to stress: A development of Stress Reduction Theory in the context of contemporary high-density city. Journal of Environmental Psychology 84, 101883. doi:10.1016/j.jenvp.2022.101883.

Luo, W., Chen, C., Li, H., Hou, Y., 2024. How do residential open spaces influence the older adults' emotions: A field experiment using wearable sensors. Landscape and Urban Planning 251, 105152. doi:10.1016/j.landurbplan.2024.105152.

Ma, H., Xu, Q., Zhang, Y., 2023. High or low? Exploring the restorative effects of visual levels on campus spaces using machine learning and street view imagery. Urban Forestry & Urban Greening 88, 128087. doi:10.1016/j.ufug.2023.128087.

Malekzadeh, M., Willberg, E., Torkko, J., Toivonen, T., 2025. Urban attractiveness according to ChatGPT: Contrasting AI and human insights. Computers, Environment and Urban Systems 117, 102243. doi:10.1016/j.compenvurbsys.2024.102243.

Mavros, P., J Wälti, M., Nazemi, M., Ong, C.H., Hölscher, C., 2022. A mobile EEG study on the psychophysiological effects of walking and crowding in indoor and outdoor urban environments. Scientific Reports 12, 18476. doi:10.1038/s41598-022-20649-y.

McGinn, A.P., Evenson, K.R., Herring, A.H., Huston, S.L., Rodriguez, D.A., 2007. Exploring Associations between Physical Activity and Perceived and Objective Measures of the Built Environment. Journal of Urban Health 84, 162–184. doi:10.1007/s11524-006-9136-4.

Meir, A., Oron-Gilad, T., 2020. Understanding complex traffic road scenes: The case of child-pedestrians' hazard perception. Journal of Safety Research 72, 111–126. doi:10.1016/j.jsr.2019.12.014.

Molléri, J.S., Petersen, K., Mendes, E., 2016. Survey Guidelines in Software Engineering: An Annotated Review, in: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, New York, NY, USA. pp. 1–6. doi:10.1145/2961111.2962619.

Navarrete-Hernandez, P., Luneke, A., Truffello, R., Fuentes, L., 2023. Planning for fear of crime reduction: Assessing the impact of public space regeneration on safety perceptions in deprived neighborhoods. Landscape and Urban Planning 237, 104809. doi:10.1016/j.landurbplan.2023.104809.

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, 105073. doi:10.1016/j.landurbplan.2024.105073.

Oku, H., Fukamachi, K., 2006. The differences in scenic perception of forest visitors through their attributes and recreational activity. Landscape and Urban Planning 75, 34–42. doi:10.1016/j.landurbplan.2004.10.008.

Protogerou, C., Hagger, M.S., 2020. A checklist to assess the quality of survey studies in psychology. Methods in Psychology 3, 100031. doi:10.1016/j.metip.2020.100031.

Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., Huang, X., 2023. Subjective and objective measures of streetscape perceptions: Relationships with property value in Shanghai. Cities 132, 104037. doi:10.1016/j.cities.2022.104037.

Qiu, Y., Wu, M., Huang, Q., Kang, Y., 2025. Do You Know Your Neighborhood? Integrating Street View Images and Multi-task Learning for Fine-Grained Multi-Class Neighborhood Wealthiness Perception Prediction. Cities 158, 105703. doi:10.1016/j.cities.2025.105703.

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, 104002. doi:10.1016/j.landurbplan.2020.104002.

Rea, L.M., Parker, R.A., 2014. Designing and Conducting Survey Research: A Comprehensive Guide.

Reynolds, J.H., Thompson, W.L., Russell, B., 2011. Planning for success: Identifying effective and efficient survey designs for monitoring. Biological Conservation 144, 1278–1284. doi:10.1016/j.biocon.2010.12.002.

Rui, J., Cai, C., 2025. Plausible or misleading? Evaluating the adaption of the place pulse 2.0 dataset for predicting subjective perception in Chinese urban landscapes. Habitat International 157, 103333. doi:10.1016/j.habitatint.2025.103333.

Rundle, A.G., Bader, M.D.M., Richards, C.A., Neckerman, K.M., Teitler, J.O., 2011. Using Google Street View to Audit Neighborhood Environments. American Journal of Preventive Medicine 40, 94–100. doi:10.1016/j.amepre.2010.09.034.

Salesses, P., Schechtner, K., Hidalgo, C.A., 2013. The Collaborative Image of The City: Mapping the Inequality of Urban Perception. PLoS ONE 8, e68400. doi:10.1371/journal.pone.0068400.

Salminen, E.A., Hausner, V.H., Ancin Murguzur, F.J., Engen, S., 2025. Optimizing recruitment in an online environmental PPGIS—is it worth the time and costs? International Journal of Geographical Information Science 39, 652–674. doi:10.1080/13658816.2024.2427267.

Shayestefar, M., Pazhouhanfar, M., van Oel, C., Grahn, P., 2022. Exploring the Influence of the Visual Attributes of Kaplan's Preference Matrix in the Assessment of Urban Parks: A Discrete Choice Analysis. Sustainability 14, 7357. doi:10.3390/su14127357.

Snowden, R.J., Snowden, R., Thompson, P., Troscianko, T., 2012. Basic Vision: An Introduction to Visual Perception.

von Stülpnagel, R., Binnig, N., 2022. How safe do you feel? – A large-scale survey concerning the subjective safety associated with different kinds of cycling lanes. Accident Analysis & Prevention 167, 106577. doi:10.1016/j.aap.2022.106577.

Suppakittpaisarn, P., Chang, C.Y., Deal, B., Larsen, L., Sullivan, W.C., 2020. Does vegetation density and perceptions predict green stormwater infrastructure preference? Urban Forestry & Urban Greening 55, 126842. doi:10.1016/j.ufug.2020.126842.

Suppakittpaisarn, P., Larsen, L., Sullivan, W.C., 2019. Preferences for green infrastructure and green stormwater infrastructure in urban landscapes: Differences between designers and laypeople. Urban Forestry & Urban Greening 43, 126378. doi:10.1016/j.ufug.2019.126378.

Taherdoost, H., 2016. How to design and create an effective survey/questionnaire; a step by step guide. International Journal of Academic Research in Management (IJARM) 5, 37–41.

Tang, J., Long, Y., 2019. Measuring visual quality of street space and its temporal variation: Methodology and its application in the Hutong area in Beijing. Landscape and Urban Planning 191, 103436. doi:10.1016/j.landurbplan.2018.09.015.

Thurstone, L.L., 1974. A Law of Comparative Judgment, in: Scaling.

Wade, N., Swanston, M., 2012. Visual Perception: An Introduction, 3rd Edition. London. doi:10.4324/9780203082263.

Wang, R., Liu, Y., Lu, Y., Zhang, J., Liu, P., Yao, Y., Grekousis, G., 2019a. Perceptions of built environment and health outcomes for older Chinese in Beijing: A big data approach with street view images and deep learning technique. Computers, Environment and Urban Systems 78, 101386. doi:10.1016/j.compenvurbsys.2019.101386.

Wang, R., Zhao, J., Meitner, M.J., Hu, Y., Xu, X., 2019b. Characteristics of urban green spaces in relation to aesthetic preference and stress recovery. Urban Forestry & Urban Greening 41, 6–13. doi:10.1016/j.ufug.2019.03.005.

Wang, Y., Zeng, Z., Li, Q., Deng, Y., 2022. A Complete Reinforcement-Learning-Based Framework for Urban-Safety Perception. ISPRS International Journal of Geo-Information 11, 465. doi:10.3390/ijgi11090465.

Wang, Z., Ito, K., Biljecki, F., 2024. Assessing the equity and evolution of urban visual perceptual quality with time series street view imagery. Cities 145, 104704. doi:10.1016/j.cities.2023.104704.

Wu, H., Leung, S.O., 2017. Can Likert Scales be Treated as Interval Scales?—A Simulation Study. Journal of Social Service Research 43, 527–532. doi:10.1080/01488376.2017.1329775.

Wu, Y., Liu, Q., Hang, T., Yang, Y., Wang, Y., Cao, L., 2024. Integrating restorative perception into urban street planning: A framework using street view images, deep learning, and space syntax. Cities 147, 104791. doi:10.1016/j.cities.2024.104791.

Yan, Y., Feng, C.C., Huang, W., Fan, H., Wang, Y.C., Zipf, A., 2020. Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience. International Journal of Geographical Information Science 34, 1765–1791. doi:10.1080/13658816.2020.1730848.

Yang, N., Deng, Z., Hu, F., Chao, Y., Wan, L., Guan, Q., Wei, Z., 2024. Urban perception by using eye movement data on street view images. Transactions in GIS n/a. doi:10.1111/tgis.13172.

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

Yao, Y., Zhu, X., Xu, Y., Yang, H., Wu, X., Li, Y., Zhang, Y., 2012. Assessing the visual quality of green landscaping in rural residential areas: the case of Changzhou, China. Environmental Monitoring and Assessment 184, 951–967. doi:10.1007/s10661-011-2012-z.

Ye, Y., Zeng, W., Shen, Q., Zhang, X., Lu, Y., 2019. The visual quality of streets: A human-centred continuous measurement based on machine learning algorithms and street view images. Environment and Planning B: Urban Analytics and City Science 46, 1439–1457. doi:10.1177/2399808319828734.

Zeng, Q., Gong, Z., Wu, S., Zhuang, C., Li, S., 2024. Measuring cyclists' subjective perceptions of the street riding environment using K-means SMOTE-RF model and street view imagery. International Journal of Applied Earth Observation and Geoinformation 128, 103739. doi:10.1016/j.jag.2024.103739.

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. doi:10.1016/j.landurbplan.2018.08.020.

Zhang, F., Zu, J., Hu, M., Zhu, D., Kang, Y., Gao, S., Zhang, Y., Huang, Z., 2020. Uncovering inconspicuous places using social media check-ins and street view images. Computers, Environment and Urban Systems 81, 101478. doi:10.1016/j.compenvurbsys.2020.101478.

Zhang, L., Tan, P.Y., Richards, D., 2021. Relative importance of quantitative and qualitative aspects of urban green spaces in promoting health. Landscape and Urban Planning 213, 104131. doi:10.1016/j.landurbplan.2021.104131.

Zhao, T., Liang, X., Tu, W., Huang, Z., Biljecki, F., 2023. Sensing urban soundscapes from street view imagery. Computers, Environment and Urban Systems 99, 101915. doi:10.1016/j.compenvurbsys.2022.101915.

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, 104421. doi:10.1016/j.jag.2025.104421.