Universität Zürich[UZH]

Geographisches Institut

Seeing Street Color, Sensing Street Mood

GEO 511 Master's Thesis

Author: Nian Li, 24-743-759

Supervised by: Prof. Dr. Sara Irina Fabrikant

Faculty representative: Prof. Dr. Sara Irina Fabrikant

26.03.2026

Universität Zürich[UZH]

Department of Geography

Seeing Street Color, Sensing Street Mood

GEO 511 Master's Thesis

Author: Nian Li, 24-743-759

Supervised by: Prof. Dr. Sara Irina Fabrikant

Faculty representative: Prof. Dr. Sara Irina Fabrikant

26.03.2026

Abstract

Urban streets influence how people feel and affect their well-being. With street view images and deep learning, citywide perception maps are now possible. However, the effect of compositional street color features is still unclear and may depend on context. This thesis checks if color affects emotional perceptions of streets using a scalable method.

Zurich is the main case study. Streets were picked from the OSMnx drive network, spaced regularly, and matched to Google Street View images using each location's unique panorama identifier (pano_id). For each pano_id, both forward and backward views were used. Four perceptions—Beautiful, Safe, Boring, and Depressing—were predicted with a Vision Transformer (ViT) using the Place Pulse 2.0 dataset. Then, Mask2Former scene-sharing controls were used. After correcting for white balance, colors were converted to CIELAB and summed up using color complexity (C_C) and color harmony (C_H).

Perception scores and color metrics both show structured spatial variation across the city, and the two exhibit a degree of association. After controlling for scene content, the addition of color variables further improves the fit of hierarchical multiple regression models, indicating that color retains a net explanatory effect beyond semantic composition. Results from local models further reveal that these effects vary across space. Building on these findings, a short test in Geneva reproduces the main directional patterns in another city, thereby supporting the framework's transferability and scalability.

Keywords: street color, urban perception, SVI, deep learning, Place Pulse, color complexity, color harmony, semantic segmentation, spatial distribution, GWR, Zurich, multiple regression

Acknowledgement

I would like to sincerely thank my supervisor, Prof. Sara Irina Fabrikant. I am deeply grateful for the time she dedicated to countless meetings, as well as for her thoughtful guidance and clear, detailed feedback on both the conceptual and technical parts of this project. Her support was essential in helping me work through many challenges.

On a personal note, I would also like to thank my parents and my girlfriend for their constant love and support throughout my studies and during the completion of my master's thesis at the University of Zurich.

Finally, I am grateful for the friends I have met along the way. Over the past few years, their friendship has made my academic journey far more enriching and meaningful.

Contents

1 Introduction

1.1 Motivation

As urbanization accelerates, urban sustainability and residents' quality of life have received increasing attention. Cities now accommodate more than 50% of the global population and produce roughly 70%-80% of global economic activity. While this concentration creates major opportunities, it also introduces major challenges for sustainable urban development (Schneider et al., 2010; Lu et al., 2018). A growing body of research indicates that higher-density urban living may be associated with mental health risks (Beenackers et al., 2024). These risks do not arise from urbanization in the abstract; rather, they often develop through repeated, everyday exposure to particular urban environments. Within this framing, streets and surrounding neighborhoods represent key spatial units shaping lived experience and well-being. Streets are complex settings where multiple components of the built environment converge. From an urban design perspective, streets also operate as "places" that enable daily activity, social interaction, and gathering (Ma et al., 2021). As fundamental elements of cities, streets go beyond facilitating movement and urban functions. They shape the city's facade, form, and image, and they strongly influence how residents evaluate urban spatial quality. The visual character of streets can also reflect political, economic, and cultural conditions, becoming part of a city's landscape and public image. Evidence across disciplines suggests that well-designed street environments can support well-being and quality of life and may help mitigate certain health risks (Van Den Bosch & Ode Sang, 2017; Brownson et al., 2009; Carmona et al., 2018; Tang & Long, 2019).

Urbanization transforms rural areas into urban landscapes and puts pressure on natural ecosystems. As interest in everyday urban experience has grown, the research field of urban perception has expanded as well (Matsuoka and Kaplan, 2008). Urban perception involves preferences and evaluations people form when viewing urban scenes (Porzi et al., 2015). People judge streets, parks, and buildings—ways the built environment shapes mental health. At the center of this work stands environmental psychology, which connects physical features with behavioral and psychological responses (Ewing and Handy, 2009). More specifically, perceptions are shaped by objective qualities of urban environments and, in turn, influence psychological well-being and spatial behavior (Nasar, 1989). Vision is often considered the most important way people perceive cities, accounting for about 83% of perception (Alexander, 2004). Within visual experience, color stands out: people often notice the color of streets or buildings before attending to their detailed shapes (Thurmann-Moe, 2017). Earlier studies have also found that color cues partly shape first aesthetic impressions.

Against this backdrop, urban color provides a useful entry point for studying urban perception. Research in color psychology suggests that certain colors can influence mood and emotional states (Elliot and Maier, 2014; Manav, 2017). For example, redder tones are often associated with liveliness, while higher saturation or brightness is more likely to evoke excitement. Color combinations also matter (Solli and Lenz, 2011; Wilms and Oberfeld, 2018). In particular, color harmony can shape visual aesthetics and influence pleasure. In urban studies, basic color attributes—hue, saturation, and brightness—have been linked to perception outcomes (Hu et al., 2023), and researchers have proposed indicators for compositional color features, such as color harmony, color diversity, and color richness (Jiang et al., 2022; Chen et al., 2022; Zhai et al., 2023). Many studies also treat color as a major component of perceived urban appearance and an element of urban identity. Research on color is well developed in areas such as marketing and tourism, and planning applications have often focused on its use. However, residents usually form impressions within the full streetscape and its spatial context, rather than from single buildings in isolation. Existing evidence suggests that urban color can shape emotional experience directly or indirectly, for example, by influencing daily activity patterns, supporting social interaction, and increasing exposure to green space (Subiza-Pérez et al., 2020). Even so, street-scale evidence remains limited on how compositional color features relate to urban perception, what mechanisms may be involved, and which other color characteristics have received less attention. More systematic work is needed to address these gaps.

Traditionally, human perception has been assessed using interviews, questionnaire surveys, and environmental audits (Duan et al., 2018; Moore et al., 2006; Griew et al., 2013). These approaches usually depend on open-ended responses or ratings based on rank or Likert-type scales (Bonaiuto, 2004). While they can inform scientific decision-making, they also have clear limitations. They are expensive and time-consuming, difficult to maintain across large urban areas, and frequently limited by sample size. When samples are not representative, response bias can arise and reduce the consistency and transferability of the findings (Liu et al., 2019). In practice, many studies remain feasible mainly in well-resourced contexts or under narrow sampling settings, which can constrain spatial coverage (Qiu et al., 2021). As a result, high spatial resolution is often achieved only for localized areas. For researchers, planners, and city officials, a principal challenge therefore remains: how to capture people's ratings and preferences across dense urban regions at scale.

Big data has changed how we study cities. Each day, people share many geotagged images publicly, giving researchers a broader view of urban environments. Street-level photos help us understand the built environment in new ways, allowing large-scale analysis of streets and urban spaces (He et al., 2023). Deep learning has advanced computer vision, enabling us to efficiently find and measure many urban features at a lower cost. The large, open-access image collections enable mapping and analysis of visual elements such as urban color (Dai, 2021). Measuring street-level quality with images is important both in theory and practice, as it addresses problems with older methods and supports efforts to improve quality of life and sustainable development (Rui & Cheng, 2023). Looking at large sets of street images lets researchers and planners turn lived experiences into data and reveal citywide patterns that traditional tools might miss, guiding better policies and decisions.

Although perception is mostly subjective, research shows that street imagery and deep learning can help map how people perceive urban areas (Dubey et al., 2016; Zhang et al., 2018; Yao et al., 2019). Street imagery provides an eye-level view of the street, similar to how residents experience their surroundings every day. This approach can provide more reliable measurements (Long and Ye, 2019). Besides, many people find it hard to state their absolute preferences, but making comparisons between options is easier and often more consistent (Nasar, 1989). Comparing options also helps people learn about subjective qualities like style and taste (Dubey et al., 2016). Furthermore, people often agree on how they see many physical features in landscapes (Bonaiuto, 2004). For example, there is a common preference for natural elements, especially plants and water (Smardon, 1988). Based on these findings, deep learning models that use street images and large datasets can estimate how people perceive cityscapes with reasonable accuracy across large areas.

1.2 Research questions

In summary, building on prior work, this study uses large-scale street imagery and deep learning to systematically assess how street color relates to urban perception. The findings provide evidence on this association and offer practical insights for sustainable color planning and spatial design.

Following this objective, the research questions of this master's thesis are:

(1) What spatial patterns characterize human emotional perceptions of street environments?

(2) What spatial distribution patterns do streetscape color complexity and color harmony exhibit?

(3) How do color complexity and color harmony influence human emotional perceptions of streetscapes?

1.3 Thesis Structure

The thesis is organized into six chapters. It moves from the theoretical background and methods to the results and validation:

- 1 Introduction: This chapter introduces the background, research gaps and motivation, states the three research questions, and explains why it is useful to quantify streetscape color and relate it to emotional perception.
- 2 Literature Review: This chapter reviews prior work on street-level urban perception, urban color theory and measurement, and the use of street view imagery and deep learning in urban research. It also identifies the main gaps addressed in this thesis.
- 3 Data and Methods: This chapter describes the study area and data sources and presents the full workflow.
- 4 Results and Analysis: This chapter reports the main results for Zurich. It presents spatial patterns in perception scores and color metrics and summarizes the regression results, including how relationships vary across space. It also includes a cross-city validation in Geneva to test robustness and transferability.
- 5 Discussion: This chapter interprets the findings in relation to the three research questions, outlines key limitations, and suggests directions for future work.
- 6 Conclusion: The final chapter summarizes the main contributions and key takeaways of the thesis.

2 Literature Review

2.1 Urban Streetscape Research

Streets are core spatial units that support many urban functions and have long been central to urban studies. In the 1960s, concerns about spatial segregation, declining urban quality, and the decline of street life led scholars to rethink what streets mean for city life. Important figures such as Jane Jacobs (Row and Jacobs, 1962) and Henri Lefebvre pointed out the social and economic consequences of street space. Since then, streets have increasingly been seen as key links between the physical form of cities and everyday activity. This evolving understanding has kept street form and function at the center of urban research.

Building on this continued attention to street form and function, recent years have seen street view imagery (SVI) become common in urban research. Its potential is increasingly clear. SVI captures streets from near pedestrian eye level and does so continuously along the road network. This is closer to everyday street experience than aerial or satellite views. Broad coverage and large data volumes matter here, and so does the relatively low cost of acquiring SVI at scale. Together, these features have widened what can be examined digitally at the city level. SVI has been used for real estate valuation (Law et al., 2019; Tran, 2024), demographic analysis (Gebru et al., 2017), therapeutic landscape assessment (Yin et al., 2020), and pedestrian flow estimation (Chen et al., 2020). It has also been applied to explaining and predicting crime rates (Xie et al., 2022; Kadiyam, 2021; Yue et al., 2024), urban accessibility analysis (Najafizadeh & Froehlich, 2018; Weld et al., 2019), and measuring the green view index (Li et al., 2015; Li et al., 2018; Li, 2021; Zhang & Zeng, 2024). Overall, this work has helped build new data foundations and pragmatic methods for studying urban structure and function.

2.2 Urban Perception and Measurement Approaches

2.2.1 Urban Perception

Urban perception is generally considered the outcome of people's interactions with their environment (Ewing and Handy, 2009). Many researchers examine how features of the built environment influence experience. Nasar (1994) found that moderate complexity and organization produce more pleasure, while high levels create greater excitement. Weber et al. (2008) found that consistency and symmetry shape aesthetic judgments.

Regarding physical elements, Zhang et al. (2018) found that different scene components shape perception. Natural features like trees or grass make places appear more beautiful, while artificial elements such as walls are linked to lower ratings, especially on safety. Quercia et al. (2014) found natural elements most strongly associate with beauty, calm, and happiness, while built features more often evoke negativity. Hu et al. (2023) found red, orange, and blue draw attention, while black conveys safety. Low saturation and high brightness make places more attractive; dull, dark colors have the opposite effect. Although design and physical elements are widely studied, urban color receives less attention, and research often focuses on basic qualities rather than color combinations.

2.2.2 Measuring Urban Perception

Many studies show our surroundings affect our well-being and behavior (Frank & Engelke, 2001; Hong & Jeon, 2015; Ulrich, 1979). Over time, experts in geography, urban planning, environmental psychology, and neuroscience have explored how our environment affects our thinking and feelings (Kaplan & Kaplan, 1989; Lynch, 1960; Nasar, 1997; Tuan, 1977). In *Space and Place: The Perspective of Experience*, Tuan was among the first to argue that "space" becomes "place" when people form emotional bonds with it.

Urban perception is subjective, complex, and context-reliant (Wei et al., 2022). It also reflects shared ideas that develop through living in cities. Because of this, urban perception matters to researchers, planners, and decision-makers who want to improve cities and support well-being (Ji et al., 2021). Studies in this field help explain how people react to the built environment and can guide planning decisions.

To measure urban perception, Quercia et al. (2014) mapped three qualities—beauty, tranquility, and happiness—across London streets, using ideas from classic urban studies and public discussions. Wartmann et al. (2021) studied aesthetic perception in Switzerland with four abstract statements. Salesses et al. (2013) examined safety, uniqueness, and social class in cities in the United States and Austria, while Ordonez and Berg (2014) conducted a similar study in Baltimore and Chicago. Many of these studies used the Place Pulse 1.0 dataset, which collected public ratings of many urban images of different qualities. When Place Pulse 2.0 was released, this research expanded. For example, Zhang et al. (2018) studied six perception qualities in Beijing and Shanghai. Wei et al. (2022) focused on four qualities and reversed one negative scale item, and Wang et al. (2024) conducted a detailed multidimensional study in Singapore and New York.

2.3 Urban Color and Measurement Approaches

2.3.1 Urban Color

Research in color psychology shows that color affects what people notice and how they feel in a scene (Elliot and Maier, 2014; Manav, 2017). Tied to specific emotions, individual colors can have unique effects (Solli and Lenz, 2011), and studies have examined hue, saturation, and brightness (Elliot and Maier, 2014). To illustrate, Goethe (1970) suggested that some hues elicit positive feelings, such as vitality and ambition; others, however, may prompt unease or anxiety. Wilms and Oberfeld (2018) similarly found that higher saturation or brightness often leads to pleasant or excited emotions.

Color effects go beyond single hues. People also react to multi-color palettes and images with many colors (Solli and Lenz, 2011). In urban studies, color shapes how we see places and affects both our emotions and thoughts. However, many studies still focus on the main content of urban scenes and treat color as less important (Lee, 2023). This is a problem because color is processed early in our vision and is often noticed before shapes. Even at low image resolution, people can still recognize broad color cues and outlines. Because of this, color is central to how we experience spaces and remains important in spatial design and urban research.

2.3.2 Measuring Urban Color

A range of approaches has been used to identify and describe urban color. For example, Zhang et al. (2021) extracted dominant facade colors for street spaces. Zhai et al. (2023) identified facade color combinations based on seven primary colors, and Zhong et al. (2021) calculated the relative weights of colors within those combinations. Many studies represent color in HSV or in the Munsell system (Zhai et al., 2023; Zhong et al., 2021) and describe it using hue, saturation, and brightness. Once colors are identified, researchers measure color features in different ways, and color harmony has received the most attention. Chen et al. (2022) and Jiang et al. (2022) assessed harmony by comparing differences between image color histograms. Zhai et al. (2023) looked at color-pair relationships within a single composition from an aesthetic perspective. Yang et al. (2024) developed a more well-defined approach based on the color wheel and grouped harmony into three types. Beyond harmony, color complexity is commonly measured using fractal-dimension methods (Zhai et al., 2023), while color richness is commonly captured with image-based metrics (Jiang et al., 2022).

Despite the flexibility of color as a design element in complex urban settings (Boeri, 2017), empirical evidence on how outdoor urban color affects residents' emotional perception remains limited. While street-level imagery has been widely used to quantify street and building characteristics in large cities (Yang et al. 2024), studies linking urban color environments to perception in Swiss cities remain relatively rare.

2.4 Street View Imagery and Deep Learning

Traditional studies of urban color focus on how buildings and streets appear based on materials and paint. These studies usually rely on field surveys using manual color charts or photo inspections (Zhang et al., 2021). Urban perception, or how people judge the built environment, is usually measured with interviews and questionnaires (Duan et al., 2018). These methods are time-consuming, expensive, and hard to scale. Small sample sizes can also cause bias and weaken results (Liu et al., 2019). In contrast, street view imagery (SVI) consists of panoramic photos collected along city streets. This approach gives a near-eye-level view of urban areas (Chen and Biljecki, 2023). As coverage expands, SVI creates new opportunities for large-scale urban analysis (Zhang et al., 2018).

Recent studies have increasingly used SVI to study urban perception. Common topics include how safe streets feel, their suitability for play, and how people view streetscapes. For instance, Ordonez and Berg (2014) developed models using image features and support vector regression. They trained these models on Place Pulse 1.0 to predict how safe streets in the United States appear to be. Porzi et al. (2015) identified visual cues that help explain why some streets feel safer. Meng et al. (2024) used SVI and deep learning to examine how people perceive residential areas. Cheng et al. (2017) found that four indicators—salient region saturation, visual entropy, green view index, and sky openness—capture key visual features of streetscapes.

Deep learning, an important part of machine learning, can discover patterns in big datasets. It has been used to simulate aspects of how people perceive images (Wang et al., 2024). When paired with SVI, it offers scalable methods for measuring both streetscape color and urban perception. For color, techniques such as semantic segmentation and K-means clustering are often used to extract color information from SVI (Zhai et al., 2023). These methods help create indicators of streetscape color (Yang et al., 2024; Zhai et al., 2023). For perception, deep learning models trained on benchmark datasets have been used to score many SVIs from different cities and categories (Wang et al., 2024; Wei et al., 2022; Zhang et al., 2018).

Overall, there is still little evidence on how street color affects urban perception, especially at the street level. Street color is a promising but often overlooked research area. To address this, our study uses SVI and deep learning to measure streetscape color features and perceived urban quality in Zurich. We employ a consistent process. We then use regression models to test how these color features relate to perception outcomes and whether these links hold after accounting for other visual factors. The aim is to give clearer evidence to support more targeted street color planning and street-level design.

3 Data and Methods

3.1 Study area overview

This study uses Zurich, Switzerland, as its case area. Zurich is the country's largest city and a key hub for services and daily activity. Its stable governance and strong reporting make it a practical setting for research on urban perception and a useful example of a dense European city.

Zurich's physical context enhances its value as a case study. The city sits at the north end of Lake Zurich, where the Limmat River flows out. Water, sloped terrain, and old neighborhood patterns shape the city's form. Streets, therefore, vary in layout, scale, and sightlines. This creates clear differences in what people see and how streets feel. These contrasts support analysis linking how streets are perceived to street-level environmental features. Zurich also has a stable internal administrative division, which makes comparisons easier and supports later spatial statistics and mapping (Figure 1).

Zurich is also well-suited from a data-and-methods standpoint. A reproducible workflow can connect street view imagery with spatial analysis, and Zurich provides strong open urban data for this purpose. The city's open data portal includes many datasets that can be linked to street segments and used for mapping, basic checks, and simple validation. Street view coverage is also extensive, which helps combine imagery with deep learning to measure urban perception and visual features.

Finally, most street-view and deep-learning studies of urban perception focus on North American and Chinese cities, while Swiss cities are rarely discussed, especially regarding street color. Studying Zurich helps fill this gap, as the city offers strong data access and is well-suited to a street-view pipeline. This, in turn, expands the geographic scope of the literature and adds quantitative evidence to inform urban design and street-quality assessment.

Figure 1. Overview of Zurich city

3.2 Data Sources

3.2.1 Base map data

(1) **Urban district data**

The urban district layer used in this study is from the “Quartiere (basierend auf Vermessungsbezirken)” dataset, released by Stadt Zürich, Geomatik + Vermessung via GeoServer. Based on official survey data and provided in the CH1903+ / LV95 system, this dataset offered Verwaltungsquartiere—Zürich's official administrative districts—as explicit spatial analysis units. Preparation in QGIS included checking the coordinate system, applying spatial filtering, and verifying topology to secure alignment with the street-view data.

(2) **Administrative boundary data**

The administrative boundaries used in this study come from the "Gemeindegrenzen (OGD)" dataset on the official Open Data Kanton Zürich portal. The dataset is published as Open Government Data (OGD), which supports its official status and reliability. For this study, the boundary of the City of Zurich was extracted as the study area. The layer was then prepared in QGIS. This included checking and harmonizing the coordinate reference system (CH1903+ / LV95), clipping to the study extent, and running basic topology and integrity checks. These steps helped ensure clean overlays with the street-view data and the other spatial datasets used in later analyses.

(3) **Basemap data**

The basemap comes from the "Basiskarte Zürich Raster" dataset by Geomatik + Vermessung, Stadt Zürich (the City of Zurich's surveying and geoinformation unit). It is based on official survey data and provided as a raster GeoTIFF in the CH1903+ / LV95 coordinate reference system. In this study, the basemap is used only for cartographic display and spatial context. Before mapping, it was checked and prepared in QGIS to match the project's spatial reference. The basemap is not used in any statistical models.

(4) Road network

The street sampling framework was based on OpenStreetMap (OSM) road data. It was generated automatically using the OSMnx package in Python. To reflect how street view images are collected and to meet the aims of this study, we used OSMnx's drive network as the base layer for sampling. Here, "drive" does not mean every road type. Instead, it shows that Google Street View is mainly captured by camera-equipped vehicles on open, physically accessible roads. Sampling points, therefore, need to fall on road types that street view cars can realistically enter and travel along. Following this logic, we kept arterial roads, secondary roads, and regular urban streets that allow motor vehicles. We filtered out pedestrian-only paths, dedicated bike lanes, mixed non-motorized paths, service roads tied to parking areas or private access, and other path types where street view coverage is unlikely. This step aligns the street network with the street view capture process and helps avoid sampling locations where images cannot be retrieved.

After acquiring the base road network, we further cleaned the raw network. This made the sampling framework more stable and the analysis more consistent. By default, OSMnx returns a directed road network. Bidirectional streets appear as two opposing edges. To avoid counting the same street segment twice simply because of travel direction, we converted the directed network into an undirected one. Each segment then enters the sampling and analysis only once. This change does not alter the network geometry. It only defines analysis units that better correspond to real, physical street segments. Intersections also tend to generate many very short edge pieces. These tiny segments can add noise and redundancy in street view sampling and spatial analysis. To limit this problem, we removed road edges shorter than 20 meters. This reduces excessive local network density and improves the stability and consistency of street-scale results

Although Zurich has high-quality official road data, OSM was used as the base for street sampling. This choice was made mainly for methodological reasons. In street view studies, the key need is not an “authoritative” road layer, but a sampling frame that is reproducible, transferable, and based on clear rules. With OSMnx, street networks can be automatically generated for cities worldwide using the same criteria. This allows the same boundary definition and filtering steps to be applied across different cities. Other researchers can then reproduce the sampling procedure in new settings. This is a core requirement in street view research (Boeing, 2017). Prior work also suggests that OSM road data are generally complete and frequently updated in urban areas, especially in highly urbanized regions. They also perform well in street-level spatial analysis (Barrington-Leigh & Millard-Ball, 2017; Moradi et al., 2022). In addition, OSM's tagging system provides a transparent way to define road access. This makes street selection rules easier to interpret and justify.

After constructing and cleaning the street sampling framework, the city of Zurich's administrative boundary was added as a spatial constraint. During system harmonization (CH1903+/LV95) and feature merging, the road network was clipped to the administrative boundary in QGIS to ensure that all street elements remained strictly within the study area. The final spatial extent of the study area and its underlying road network are shown in Figure 2.

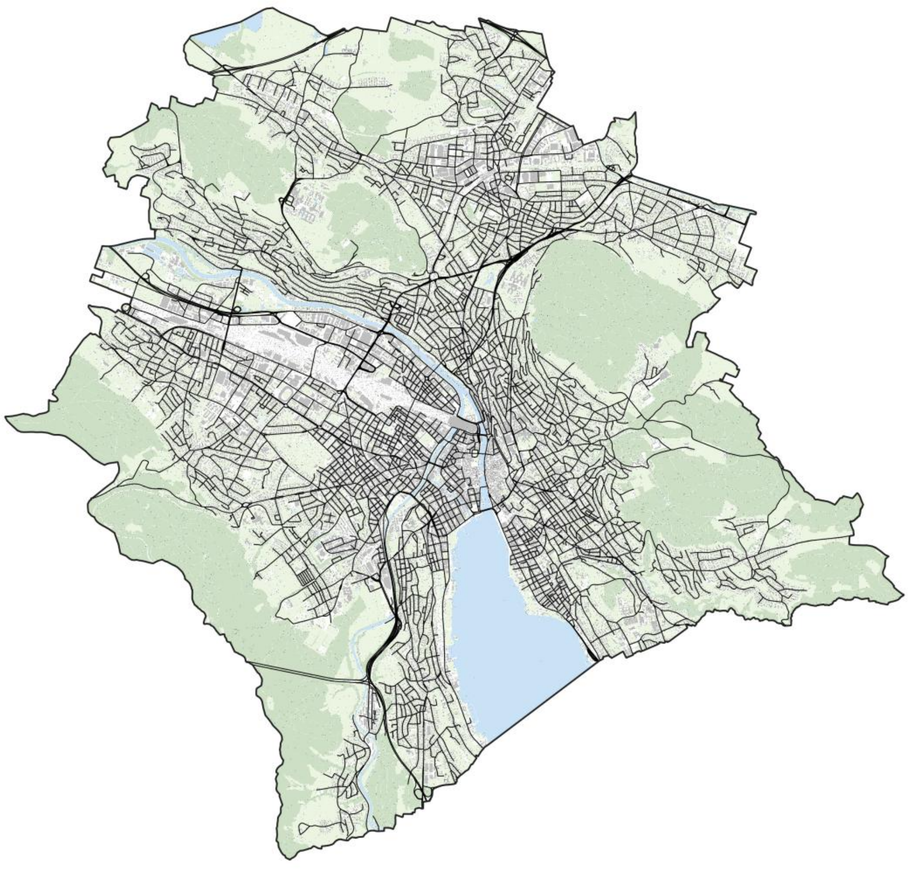


Figure 2. Overview of Zurich Road Network

3.2.2 Sampling Points

After building and cleaning the motor-vehicle-accessible road network, it became the base for street view sampling. Sampling points were placed at regular intervals along street centerlines in QGIS. This approach generated an even, reproducible set of observation locations across the study area. It also helped prevent samples from clustering along a few major roads or within a few subareas.

Before generating points, the road geometry was standardized. The raw road layer consisted of many short, segmented line features. Placing points at equal spacing on each small segment would have concentrated them near segment breaks and intersections. To mitigate this, road centerlines were merged first. MultiLineString geometries were then split into single LineString features. This created sampling units that more closely matched contiguous streets, so the chosen interval had a more consistent meaning across the network.

Sampling points were placed along the processed centerlines at a fixed 50 m interval. At the city scale, this spacing is a practical trade-off. It offers broad coverage while keeping image download and processing costs manageable (Chen et al., 2025; Qi et al., 2025). For data management, each point received a unique ID. Its coordinates and attributes were exported for street view retrieval and later analysis. Figure 3 shows the final point distribution. In total, 18,285 sampling points formed a citywide observation network across the study area.

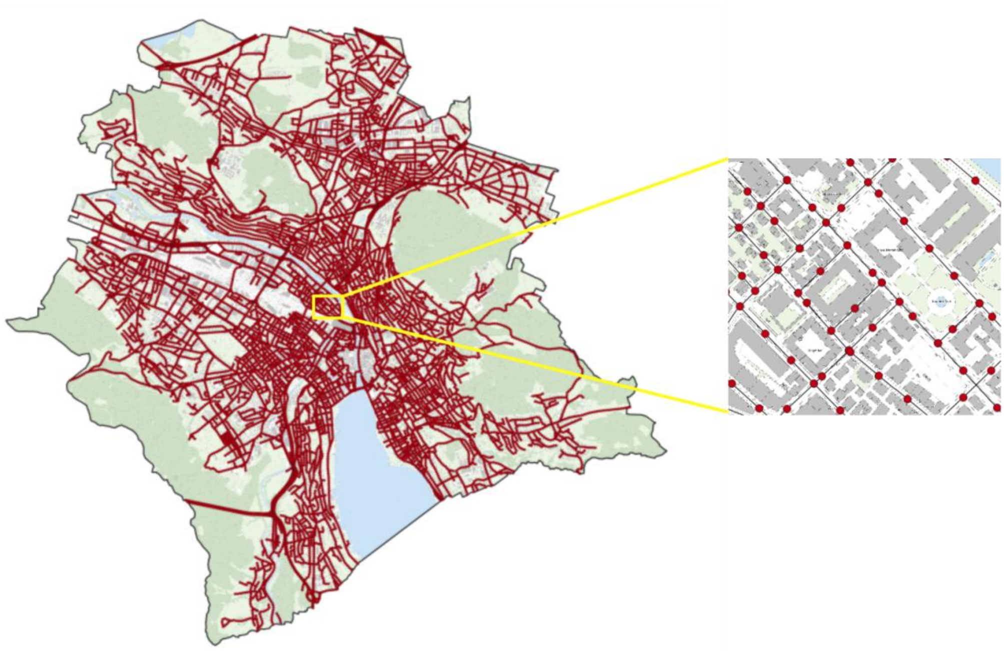


Figure 3. Spatial distribution of sampling points in Zurich

3.2.3 Google Street View Imagery

After generating sampling points, street view images were retrieved through the Google Street View (GSV) API. GSV was chosen as the visual data source because it is widely used in urban research and built-environment assessment. It can also be integrated into reproducible workflows. Prior studies show that GSV supports "virtual audits". These audits can be completed from the office and often yield results broadly comparable to on-site field audits. They also require far less time and cost. For instance, Rundle et al. (2011) and Wilson et al. (2012) report that GSV-based observations capture key streetscape features and support systematic neighborhood-scale assessment. Curtis et al. (2013) and Odgers et al. (2012) emphasize the utility and consistency of street view imagery for urban social observation and environmental exposure research.

GSV also provides a stable street-level viewpoint. The system typically captures images from within the roadway. This closely matches what people see as they move through the street environment. This feature makes GSV a good fit for studying "street experience" cues like color composition, overall feel, and visibility. In addition, GSV offers broad coverage and a consistent access method. This helps keep image definitions aligned within and across cities. It also offers a common base for later color-metric extraction and perception modeling.

To balance coverage and download cost, a two-stage workflow was used in this study. Stage 1 sent batch Street View Metadata requests for each sampling point. This check verifies if usable imagery is available nearby and returns the panorama ID (pano_id) and matched location. Stage 2 downloads full images only for panoramas marked as available. This avoids unnecessary downloads where there is no street view coverage. In metadata queries, each point's latitude and longitude served as the search location with a 40 m radius. This value was selected through sensitivity testing. A smaller radius lowers the hit rate, while a larger radius increases the likelihood of matching imagery from another street rather than the target segment. A 40 m radius offers a workable trade-off between coverage and alignment. To reduce interference from non-street scenes, metadata requests were restricted to outdoor scenes. This prioritizes outdoor street view imagery.

After retrieving metadata for all sampling points, only valid records were kept. Repeated pano_ids were removed because the same panorama can be returned for multiple nearby points. Without deduplication, identical panoramas would be downloaded multiple times. This would create redundant samples. After filtering and deduplication, the remaining panoramas were used for image downloads. Representative examples are shown in Figure 4.

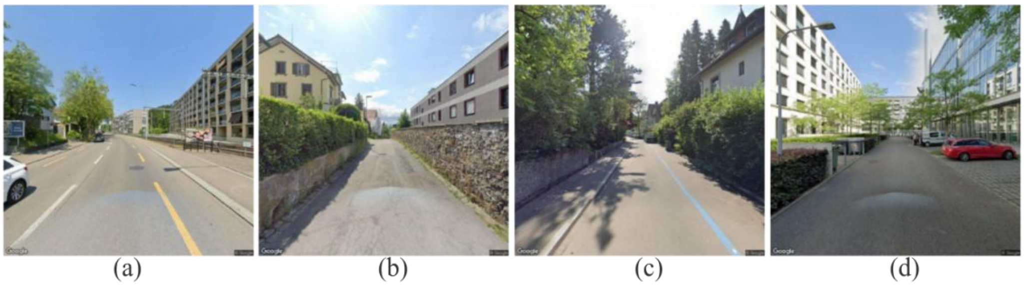


Figure 4. Examples of GSV

For downloading, we used the deduplicated pano_id list as an index and retrieved images through the street view image API. Rather than choosing random or fixed headings, we aligned the views with the street's direction to match how the street view vehicle moves. Each pano_id was matched to the closest road centerline, and we used the local road direction as the heading. We downloaded two images per pano_id: one facing forward and one facing backward (see Figure 5). This approach keeps the sample size manageable, captures both directions, and reduces noise from random angles.

Figure 5. Schematic diagram of street view image crawling

After the initial download, all images underwent a basic quality check. Records with abnormal responses, corrupted files, or clearly invalid content were removed. After filtering, 11,854 unique pano_ids were retained, corresponding to 23,700 valid street view images. These images were used as inputs for subsequent steps. All retained images were captured with the same field of view, resolution, and pitch to reduce bias related to differences in imaging settings.

3.2.4 Perception Dataset

This study uses Place Pulse 2.0 from the MIT Media Lab as the reference dataset for urban perception (Dubey et al., 2016). Place Pulse is an online crowdsourcing platform where participants compare street-level images and report how places "feel." By the end of 2016, the project had gathered over 11,169,000 pairwise comparisons from 81,630 participants. Place Pulse 2.0 has since been used widely in studies of urban perception and the built environment, and it remains a key empirical resource in this area.

In the original task design of Place Pulse 2.0, participants see two randomly selected street view images from the same city and choose which one better matches a given perceptual attribute (Figure 6). The dataset covers six dimensions: safe, beautiful, lively, wealthy, depressing, and boring. Pairwise choices help avoid many of the issues that come with direct numeric ratings, since people do not use rating scales in the same way. This design tends to produce perception measures that are more stable and more comparable across participants.

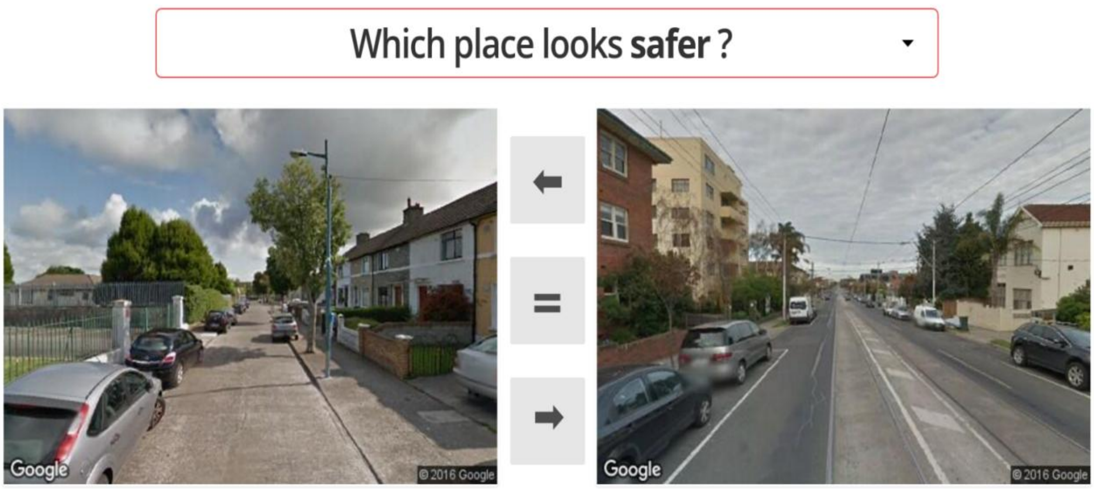


Figure 6. Interface of the MIT Place Pulse 2.0 data collection platform. Participants were shown two street view images and asked which one better matched a given perceptual attribute. The dataset was constructed from a large number of these pairwise comparisons (Dubey et al., 2016).

In terms of score computation, Place Pulse 2.0 derives continuous perception intensity indicators from pairwise comparison results. For a given perception dimension, let w_i, l_i, and t_i denote the number of times image i is selected, not selected, or judged as equal in comparisons, respectively. The basic win ratio of image W_i can be expressed as:

$$W_i = \frac{w_i}{w_i + l_i + t_i}$$

However, relying solely on the win ratio may be influenced by the relative strength of comparison counterparts. Therefore, previous studies further introduced a correction term that incorporates the average win ratios of the images compared against image i, resulting in a more stable perception score (Salesses et al., 2013). For each image and perception dimension, the final score is normalized and mapped to a continuous 0-10 scale. This produces the standard Place Pulse perception score (Q-score). The Q-score offers a robust summary of how people rate the relative strength of each perceptual attribute in a given street view image. The experiment attracted volunteers from 162 countries, reflecting a high degree of geographic and cultural diversity. Existing studies have shown that the Place Pulse dataset and its perception results are not statistically dependent on participants' age, gender, or geographic location, and that differences in cultural background, income level, or ethnicity do not introduce systematic bias into the overall results (Salesses et al., 2013; Dubey et al., 2016). As a result, the dataset is considered to exhibit strong cross-regional applicability and has been widely used to analyze human perceptions of urban visual environments.

In this study, based on the research objectives and subsequent analytical needs, four perception indicators—beautiful, safe, depressing, and boring—are selected from the Place Pulse 2.0 dataset as the primary dimensions for street-level perception modeling and spatial analysis. This selection follows three principles: (1) according to Maslow's hierarchy of needs, safety and aesthetic quality constitute fundamental human requirements for environmental experience (Maslow, 1943); (2) depressing has been widely adopted in prior research as a key indicator for assessing the potential impacts of urban environments on mental health (Gong et al., 2016); and (3) selecting two positive and two negative perception dimensions facilitates a more balanced and contrastive interpretative structure in subsequent spatial analyses.

3.3 ZenSVI

To support systematic acquisition, processing, and analysis of street view imagery (SVI) at the urban scale, this study uses ZenSVI—an open-source SVI analysis toolkit—as the core technical framework. ZenSVI is an end-to-end software package developed by Ito et al. (2025) that brings key parts of street-view research into a single workflow, including image acquisition, metadata handling, image transformation, computer vision-based processing, and result visualization. This integration helps lower the technical entry barrier for street-view research and strengthens methodological transparency and reproducibility.

ZenSVI is written in Python and works smoothly with common geospatial tools such as GeoPandas and OSMnx. It can retrieve images from multiple street-view platforms and includes established computer-vision models for extracting streetscape features and producing perception-related indicators. Compared with older, ad hoc street view scripts that are often difficult to reuse, ZenSVI offers a workflow that is more modular and easier to extend (Figure 7). This makes it more straightforward to run the same pipeline across different cities and research settings. As a result, ZenSVI has been widely used in studies of street-scape measurement, urban perception, and built-environment analysis.

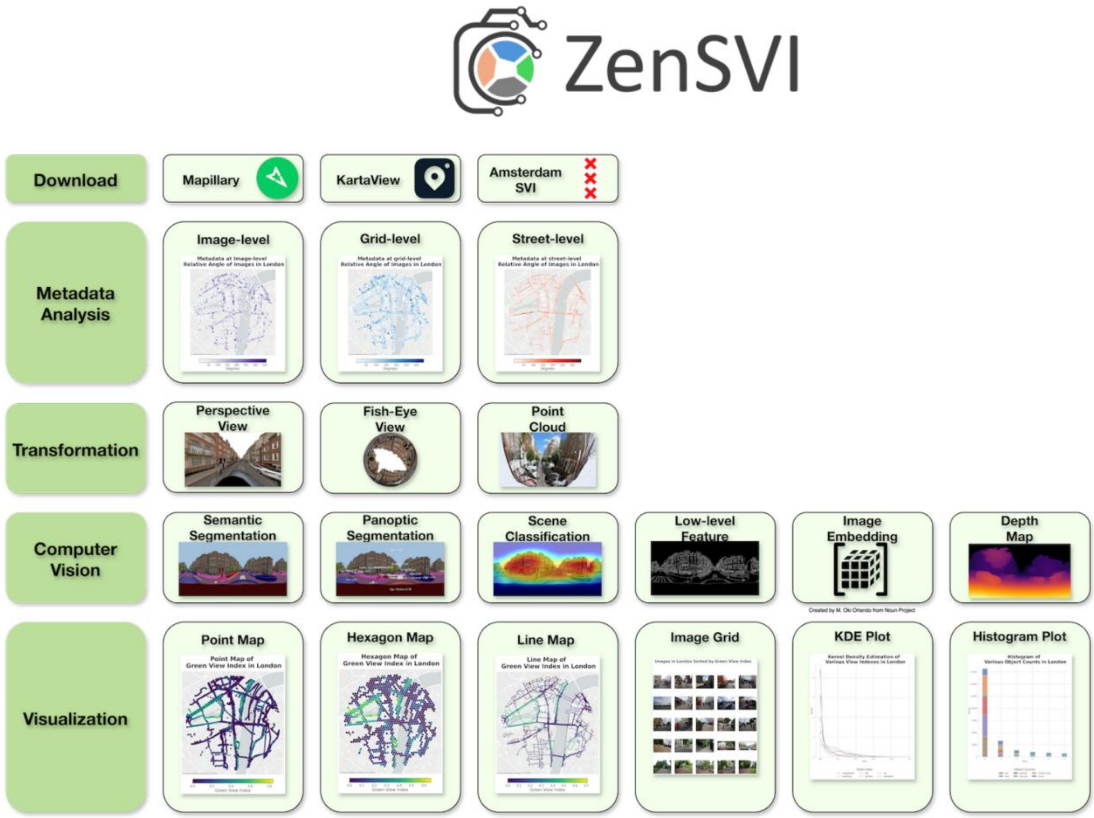


Figure 7. Overview of the framework of ZenSVI package

3.3.1 Vision Transformer (ViT) Model for Perception Scoring

This study uses the urban perception prediction model provided in the ZenSVI framework (Ito et al., 2025), specifically the ViT-based version, following the publicly released implementation by Ouyang (2023). Designed to predict subjective perceptions of urban environments from street view images, the model uses ViT-B/16 as its backbone. The network is initialized with ImageNet-pretrained weights for generic visual representations. It is then trained on the MIT Place Pulse 2.0 dataset to adapt it for urban perception tasks. With this setup, the resulting weights are ready for direct use in street-level perception inference.

As shown in Figure 8, Vision Transformer works by splitting an input image into fixed-size patches, then applying the Transformer's self-attention to learn how these patches relate to one another. This process helps capture the global structure and context across the entire image (Dosovitskiy et al., 2021). In street-view perception, human judgments typically depend on the overall scene—such as street openness, spatial continuity, and the combined layout of built and natural elements—rather than isolated local textures. Unlike convolutional neural networks, which focus on local receptive fields, ViT is better suited for tasks that require understanding the global scene and relationships across different image regions (Geirhos et al., 2019).

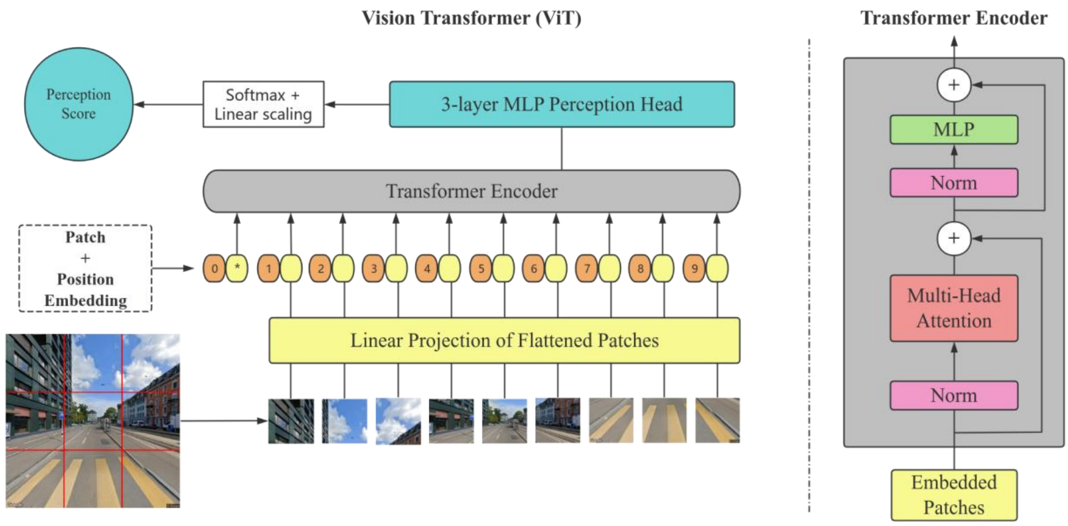


Figure 8. ViT model: the input street view image is divided into equal patches, which are converted into embedded representations. These patch embeddings are combined with positional information and a class token. The resulting sequence is then passed through multiple Transformer encoder layers to learn global image features. The final class token is used to generate a binary output, which is passed through a softmax to obtain the probability of the positive class. This probability is then linearly scaled to a score from 0 to 10, representing the predicted perception score. The Transformer encoder follows the design by Vaswani et al. (2017).

In the original implementation, the ViT perception model uses the large-scale pairwise comparison results from Place Pulse 2.0 as supervision and treats each perception dimension as a separate classification task. Six independent classifiers are trained, one for each dimension: safe, beautiful, lively, wealthy, boring, and depressing. Reported classification accuracies are 76.7% for safe, 77.1% for lively, 72.9% for wealthy, 76.9% for beautiful, 61.6% for boring, and 67.2% for depressing, suggesting that the model learns stable and consistent patterns in human perceptual judgments from street view imagery (Ouyang, 2023).

In practice, all input SVIs undergo the same preprocessing steps. Each image is then fed into the ViT classifier for the target perception dimension, producing logits that are converted to binary class probabilities using a softmax function. The probability of the positive class is taken as the perceptual intensity for that dimension and is linearly rescaled to a 0-10 range to match the commonly used Place Pulse perception scale. Because a single panorama can look different in different directions along the road axis, two images are retrieved for each pano_id—one facing forward and one facing backward—and scores are predicted separately. Mean values are then calculated at the pano-level and used as aggregated perception measures in subsequent street-scale spatial statistical analyses (Figure 9).

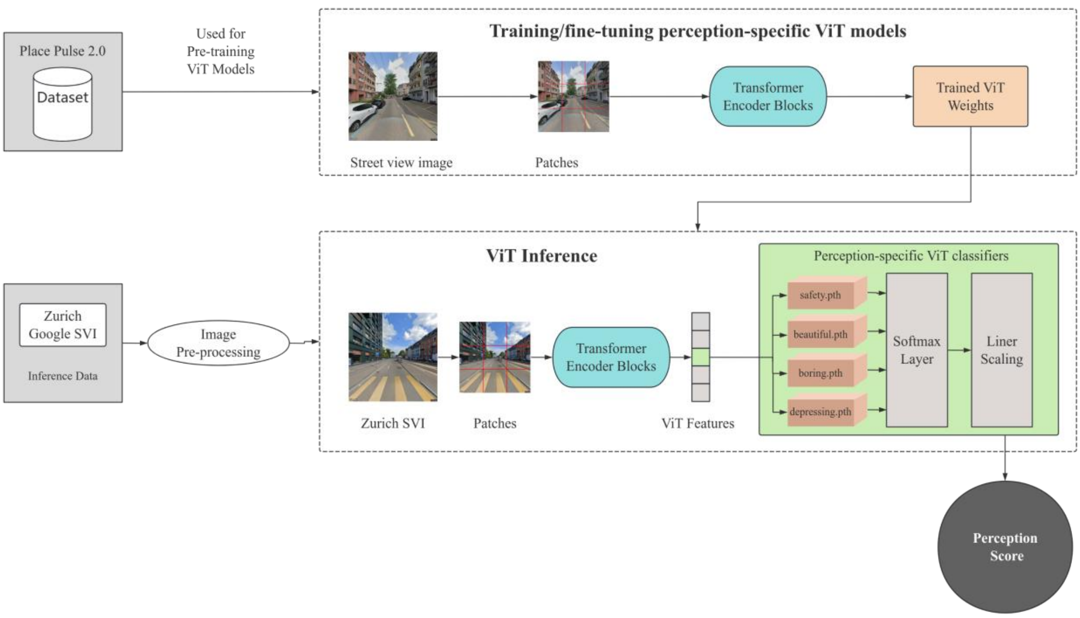


Figure 9. Workflow of ViT-based perception scoring: the perception-specific ViT models are trained/fine-tuned on the Place Pulse 2.0 dataset to obtain model weights for each perception dimension. During inference, each Zurich street-view image is pre-processed, split into patches, and fed into the Transformer encoder to extract a global feature representation. The model then loads the weights for the four perception dimensions—safe, beautiful, boring, depressing—and produces predictions separately for each dimension. The output binary logits are converted to the positive-class probability via a softmax, and the probability is linearly mapped to the 0–10 range to yield the final perception score.

Based on these considerations, ViT is employed as the core prediction model for street-view perception in this study, and the resulting perception scores serve as the primary input variables for subsequent analyses of streetscape color characteristics and spatial patterns.

3.3.2 Mask2Former Model for Semantic Segmentation

Semantic segmentation assigns each image pixel a semantic class. In urban scene analysis, it identifies elements in SVI and describes the streetscape composition and spatial distribution (Zhou et al., 2017). Elements such as buildings, vegetation, and roads shape subjective impressions (e.g., safety, beauty, boredom, and depression). Here, semantic segmentation divides scenes into components and quantifies the proportions of pixels within each component.

This study uses the Mask2Former model from the ZenSVI framework as the baseline for segmenting SVI (Figure 10). Cheng et al. (2022) introduced Mask2Former as a unified Transformer for semantic, instance, and panoptic segmentation. In ZenSVI, it acts as a standardized parsing module for reproducible pixel-level predictions and reliable comparisons across samples and cities.

Mask2Former is run in ZenSVI and uses the Mapillary Vistas label system, which covers urban elements such as buildings, roads, sky, vegetation, sidewalks, and walls. In semantic segmentation mode, the model outputs a pixel mask for each image. Two outputs are generated: (1) color-coded segmentation images for visual checks (Figure 11), and (2) CSV summaries reporting pixel proportions per category per image.

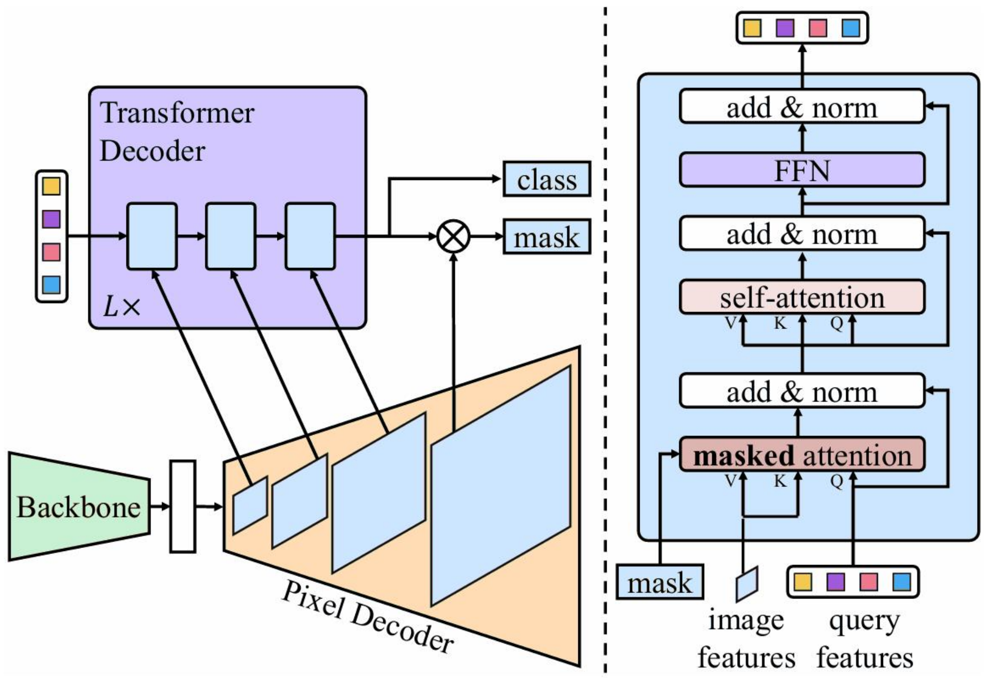


Figure 10. Mask2Former model: Mask2Former uses the same meta-architecture as MaskFormer, comprising a backbone, a pixel decoder, and a Transformer decoder. Cheng et al. (2022) introduced a redesigned Transformer decoder that replaces standard cross-attention with masked attention. To better handle small objects, the decoder takes one scale from the multi-scale features at each Transformer layer, enabling more efficient use of the high-resolution features produced by the pixel decoder. In addition, the model changes the order of self-attention and cross-attention (implemented as masked attention), allowing the query features to be learned, and removes dropout to improve computational efficiency.

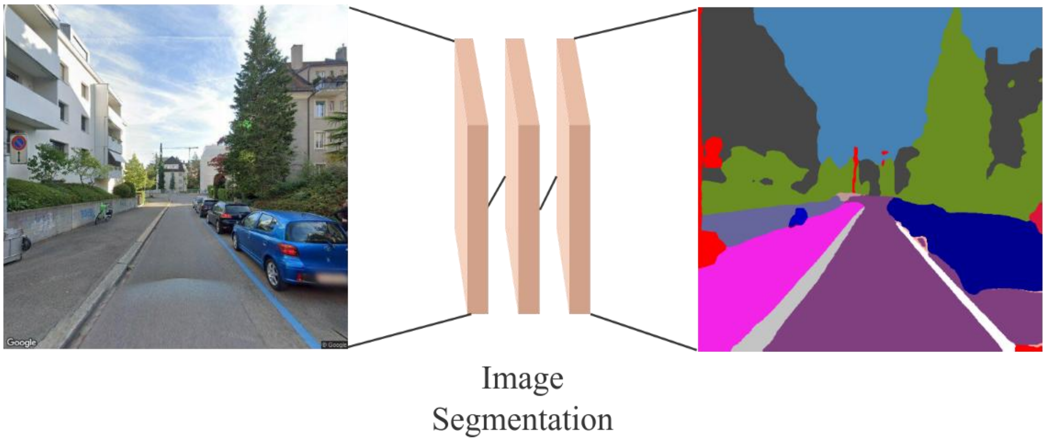


Figure 11. Example of Image Segmentation

Using these segmentation outputs, this study creates a structure-background binary mask aligned with the research objectives. As commonly done in urban scene and street perception research, the categories of buildings, roads, sidewalks, sky, walls, and vegetation are specifically grouped together as structure-background elements. Pixel-level semantic segmentation assigns all these structure-background categories to foreground pixels (value 1), while assigning all other classes a value of 0 (see Figure 12). This process produces a binary mask at the original image resolution, delivering a clear, reproducible spatial constraint for analyzing street color and distinguishing street structure from background.

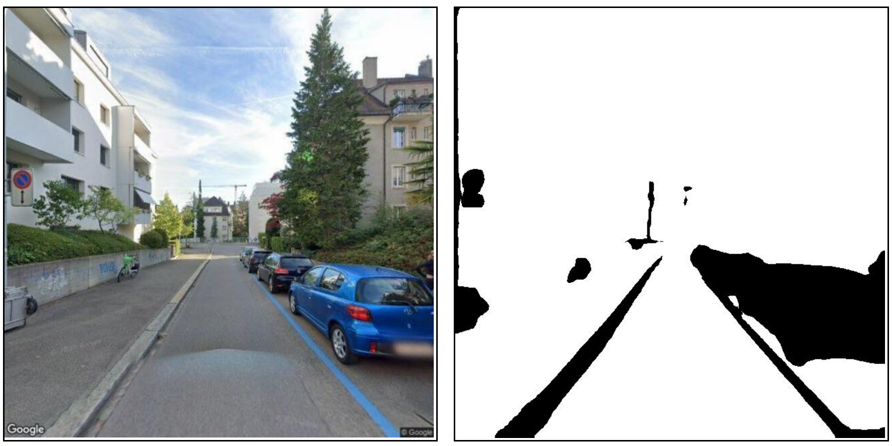


Figure 12. Example of generating a structure-background binary mask: Based on the RGB class codes in the semantic segmentation map, the pixels labeled as Building,

Road, Sidewalk, Sky, Wall, and Vegetation were merged as the foreground (1; white/255 in the mask), while all other classes were treated as the background (0; black/0). A same-size structure-background binary mask was then produced and saved as a 0/255 grayscale image.

This study stands out by including natural elements, such as the sky and vegetation, in the structure-background category. Most street-view analyses focus on man-made features, such as building facades and roads, and often leave out the sky and vegetation to reduce seasonal and weather effects. In contrast, we examine how street color composition relates to how people perceive streets. Since people notice natural features like the brightness of the sky and the color of vegetation, omitting them makes images less true to life and harder to understand how people perceive streets.

SVIs are collected passively and reflect differences in season, weather, and lighting, which is a known challenge for maintaining data consistency over time (Biljecki & Ito, 2021). Although these effects cannot be fully eliminated, this study reduces bias by carefully selecting data sources and using specific analytical methods. For example, images from Google Street View are usually captured in clear daylight with good visibility (Harada & Oki, 2025; Ibrahim et al., 2019), and collections are planned under quality conditions. Here, each pano_id is analyzed using combined predictions from two directions along the road, applying the mean to reduce sensitivity to blocked views, exposure, and lighting, and create more stable street-level data.

For the analysis, we do not focus solely on simple correlations between perception scores and color indicators. Instead, we use the proportions of different elements, like buildings, roads, vegetation, and sky, as controls in a multivariate regression. We select variables using forward selection. After accounting for the mix of scene elements and other factors, the models assess whether color features provide additional explanatory power. This helps us tell apart the effects of scene content from those of color, so that changes caused by season or weather are not mistaken for real color effects. This approach is standard in research, where visible environmental factors are controlled to focus on visual effects (Salesses et al., 2013; Dubey et al., 2016).

Although it is not possible to completely remove seasonal and weather differences from street view images, this study addresses these challenges by using standardized image sources, combining views from different angles, and applying multivariate modeling to keep the analysis focused on how people perceive streets.

3.4 Color Space Transformation

To measure street-view colors in a stable and comparable way, the images were processed before any color features were extracted. The main goal was to reduce the impact of lighting differences on color measurements. The workflow applies white-balance correction first, followed by a color-space transformation. This converts the original RGB images into a representation that is closer to human color perception and provides a consistent, interpretable base for later color-feature computation and perception analysis.

3.4.1 White-Balance Correction

Street view images (SVIs) are automatically collected, so their colors can change depending on lighting, the camera, and the surfaces in the scene. Even with a standard source like Google Street View, photos taken at different times, locations, or angles can look warmer or cooler, or exhibit a color tint. If we use the raw RGB values to measure color, lighting effects can get mixed up with actual differences in the environment. This makes it harder to compare images.

To address this, white balance correction was applied to all images before converting them to other color spaces. The Simple White Balance method in OpenCV was chosen. This method adjusts each color channel’s histogram to make the overall colors more neutral. It does not rely on knowing what is in the scene, so it works well for automated processing on large sets of images. It also balances speed and reliability.

It is important to note that white balance correction does not remove real color differences in the streetscape. Instead, it reduces color shifts caused by the camera or lighting, so later color analysis can focus on the true colors of the street environment rather than artifacts from the imaging process.

3.4.2 RGB-to-CIELAB Color Space Conversion

After white balancing, images are converted from RGB to the CIELAB ($L^*a^*b^*$) color space. This conversion is important because, unlike RGB, which depends on the device and whose values do not match how people see color differences, CIELAB is designed for perceptual uniformity. Created by the Commission Internationale de l'Éclairage (CIE), CIELAB ensures that equal distances represent roughly equal perceived color differences. As a result, it is better suited than RGB for measuring color contrasts and composite color features.

In CIELAB:

- L^* represents lightness, ranging from 0 to 100;
- a^* is the chromatic axis from green to red;
- b^* is the chromatic axis from blue to yellow.

This approach is common in visual perception research, color-difference calculations, and streetscape color analysis because it separates brightness from color. By doing so, researchers can more closely study the color composition and contrast of street scenes. In this study, we first convert white-balanced RGB images to the Lab color space with OpenCV. After conversion, values are adjusted to standard CIELAB ranges: L^* 0–100, a^* and b^* -128–127. These CIELAB values are stored as floating-point arrays for later calculations, and the 8-bit OpenCV-Lab version is kept for visualization and quality checks. This workflow assures precise analysis and easy visualization.

3.4.3 Methodological Rationale and Link to Subsequent Analyses

Converting SVIs to CIELAB under a consistent white-balance procedure serves three methodological aims:

(1) It reduces systematic variation from changing imaging conditions and improves comparability across samples;

(2) It represents streetscape color in a way that is closer to human visual perception, which better supports perception-focused analysis;

(3) It provides standardized inputs for later pixel-based metrics, including color complexity and color harmony.

With this preprocessing in place, we can extract and summarize streetscape color features more selectively. This sets up a more systematic test of how street-scape color structure relates to subjective perception.

3.5 Color Metrics Computation

SVIs have a lot of detailed color information at each pixel. To make this data easier to work with, it is simplified into a form that is easier to understand and model. First, the images undergo white balance correction and are converted to the CIELAB color space. Then, the main colors and their proportions are identified. Using this summary, two main metrics are calculated: color complexity (C_C) and color harmony (C_H). By focusing on the main colors, the process remains efficient and aligns with how people typically perceive color patterns. This approach also creates simple, reliable data that can be used in later perception models.

3.5.1 CIELAB-based K-means Clustering

While pixel-level RGB or Lab color statistics keep all the information, they are often affected by noise, outlier pixels, and redundant data. This makes it hard to compare samples and build statistical models. Previous studies have found that using a small set of dominant colors and their proportions, known as the dominant color descriptor (DCD), strikes a good balance between compressing information and maintaining visual meaningfulness. This approach is widely used in image analysis and urban color research.

Based on this idea, K-means clustering is used to group color pixels without supervision and find a limited number of dominant colors. K-means minimizes squared error within each group, thereby finding representative color centers in the color space. Unlike fixed-bin color histograms, K-means can better match the real color distribution in different SVIs. It has also been widely used to identify color combinations in SVIs (Zhai et al., 2023; Zhong et al., 2021).

For clustering, only the a^* and b^* chromatic channels in CIELAB are used, while the L^* lightness channel is left out. L^* mainly shows changes in brightness and is easily affected by lighting, shadows, and exposure during image capture. In contrast, a^* and b^* directly represent color and intensity as we see them. Studies have shown that including L^* can create false color groups because of uneven lighting, making the dominant colors less accurate for the real environment. Therefore, dominant colors are extracted only from the a^*b^* plane to focus on the true color makeup of streetscapes and to lessen the effect of lighting differences.

For the number of clusters, this study sets $K = 4$. This decision is based on earlier research in urban color and environmental perception, which found that a smaller K can reliably capture the main color structure people notice, without adding noise or making the results harder to interpret. Other studies also show that four clusters are enough to match how people group and understand colors visually (Nguyen et al., 2020). Using K=4 preserves overall color information and helps make subsequent calculations and modeling more reliable.

In implementation, cluster centers are first fitted using a random sample of pixels, after which all pixels are reassigned to their nearest centers to obtain more accurate estimates of dominant-color proportions. Each SVI is ultimately represented by K dominant color centers $\left(a_k^*, b_k^*\right)$ and their areal proportions S_k, satisfying $\sum_{k=1}^{K} S_k = 1$.

To support analysis at different semantic scopes, dominant colors were extracted in two ways. The "whole-scene" scope summarizes the color structure of the entire image. The "structure-background" scope restricts extraction to the area defined by the semantic mask, so it captures the colors of the street-space components that are most structurally relevant.

3.5.2 Color Complexity (C_C)

Color complexity (C_C) is often measured by simply counting the number of colors in an image, which can overlook the space each color occupies and its distribution, thereby reducing accuracy. To better assess street color complexity, this study uses the Herfindahl-Hirschman Index (HHI), an economic tool. HHI typically measures market concentration, but here it quantifies the concentration or diversity of color distribution by summing the squared proportions of main colors. A higher HHI means a few colors dominate (lower C_C), while a lower HHI means colors are more evenly distributed (higher C_C).

HHI was chosen for its simplicity and effectiveness in showing color complexity. Compared to other complexity measures, such as entropy or standard deviation, HHI is more practical. While entropy can also measure complexity, it is harder to calculate and explain. HHI provides a clear and concise way to analyze color in large urban studies. Accordingly, let the dominant-color proportions of an image be $S_1, S_2, \ldots, S_n$ with $\sum_{i=1}^{n} S_i = 1$. Color complexity is defined as:

$$Cc = \left| 1 - \sum_{i=1}^{n} S_i^2 \right|$$

Under this definition, when a single dominant color occupies an overwhelming share, $\sum S_i^2$ becomes large and C_C becomes small, indicating a more monotonous color structure. When multiple dominant colors are more evenly distributed, $\sum S_i^2$ becomes smaller and C_C becomes larger, indicating a richer and more complex color structure. This metric has a clear statistical interpretation, is computationally simple, and is numerically stable, making it well suited to large-scale streetscape samples. The value range of C_C is [0,1].

In this study, both whole-scene color complexity (C_C_whole) and structure-background color complexity (C_C_struct) are computed to compare differences in color diversity between the full streetscape and the structurally constrained visual region, and to provide contrasting variables for subsequent perception modeling.

3.5.3 Color Harmony (C_H)

While color complexity emphasizes the "dominant-color proportion structure," color harmony (C_H) focuses on the perceived coordination among multiple dominant colors. A substantial body of work in environmental perception and visual psychology suggests that even with the same number of colors, differences in contrast strength and color combination patterns can produce markedly different aesthetic and emotional responses. Therefore, "color count/distribution" alone is insufficient to characterize experiential differences in streetscape palettes, and a color-difference-based measure is needed to represent the degree of "harmony versus discord" (Wei et al., 2014; Li et al., 2020).

To quantify color harmony, this study adopts CIEDE2000 (ΔE_{00}) in CIELAB as the distance metric between dominant colors. Compared with earlier Euclidean distances based directly on ΔL^*, Δa^*, Δb^*, ΔE_{00} provides improved nonlinear correction and perceptual consistency, and is widely regarded as more aligned with human judgments of color difference; it is therefore more appropriate for street color evaluation tasks where perceptual difference is central (Wei et al., 2014). Given the K=4 dominant colors extracted for each image, let the CIELAB representation of dominant color i be C_i, with areal proportion S_i in the image (and $\sum_{i=1}^{K} S_i=1$). To ensure that the importance of a color pair matches its spatial contribution in the image, the product of their areal proportions is used as the weight for the color pair (i,j) (Li et al., 2020):

$$w_{ij} = S_i S_j$$

The weighted mean color difference across the dominant-color set is then defined as:

$$\bar{\Delta E}_{00} = \frac{\sum_{i<j} w_{ij} \cdot \Delta E_{00}(C_i,C_j)}{\sum_{i<j} w_{ij}}$$

This weighting prevents very small color patches from exerting a disproportionate influence on the overall color difference, making the dominant-color contrast primarily determined by color pairs that occupy larger visual shares and thus better reflecting perceived palette-level differences (Li et al., 2020). It should be noted that dominant colors are extracted via clustering in the a^*b^* plane to reduce the influence of illumination intensity on cluster centers, yet color-difference computation requires a complete (L^*, a^*, b^*) representation. To maintain consistency, the mean lightness $\bar{L}^*$ of the corresponding analysis scope (whole-scene or structure-background) is assigned to all dominant colors, forming $C_i=(\bar{L}^*,a_i^*,b_i^*)$ for ΔE_{00} computation. This treatment focuses the distance measure on hue-chroma composition while using a unified luminance baseline to avoid misinterpreting local exposure variations as structural differences among dominant colors.

To map $\Delta\bar{E}_{00}$ onto an intuitive and comparable harmony scale, it is normalized to [0,1] and defined as:

$$C_H = 1\text{-min}\left(\frac{\Delta\bar{E}_{00}}{\Delta E_{\max}}, 1\right)$$

Here, $\Delta E_{\max}$ is set to a high quantile of $\Delta\bar{E}_{00}$ over all samples as a robust scale parameter (this study uses $Q = 0.98$), reducing scale distortion induced by extreme samples and improving metric stability. Under this definition, smaller differences among dominant colors yield lower $\Delta\bar{E}_{00}$ and C_H closer to 1 (more harmonious palettes); stronger contrasts yield higher $\Delta\bar{E}_{00}$ and lower C_H. The value range of C_H is [0,1].

Consistent with the dual-scope strategy used for color complexity, color harmony is computed under both whole-scene and structure-background semantic scopes, yielding C_H_whole and C_H_struct. This setup allows a direct comparison between color coordination in the full streetscape view and in the structure-background subset. It also produces two contrasting inputs for later models, which helps test whether the color metrics add explanatory power once semantic composition is controlled.

3.6 Spatial Analysis Methods

3.6.1 Spatial Autocorrelation Analysis

Spatial autocorrelation tests whether values show a geographic pattern rather than being randomly distributed. In simple terms, it asks whether nearby locations tend to have similar values. If high values cluster near other high values (and low values near other low values), the pattern is positive spatial autocorrelation. If high and low values are mixed together, the pattern is negative. If neither pattern appears, the distribution is closer to spatial randomness. Based on this idea, global spatial autocorrelation tests are applied to the streetscape perception measures and the color metrics. The goal is to assess whether these variables exhibit an overall spatial structure and to establish a statistical baseline for subsequent local clustering analysis and spatial modeling.

Two global statistics are reported: Global Moran's I and Getis-Ord General G. They are complementary, capturing different aspects of spatial structure. Moran's I assesses whether similar values tend to be adjacent, making it useful for detecting broad clustering or dispersion in continuous variables. General G is more sensitive to the global concentration of extremes, especially the tendency for high (or low) values to cluster. Street perception and color indicators may show both kinds of structure at once—for example, a smooth center–periphery gradient and localized high-value concentrations. Using Moran's I together with General G provides a fuller description of the global spatial pattern.

The formulation of Moran's I is given as follows:

$$I = \frac{n}{\sum_{\mathrm{i}} \sum_{\mathrm{j}} w_{ij}} \cdot \frac{\sum_{\mathrm{i}} \sum_{\mathrm{j}} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where n is the number of observations, x_i is the value observed at location i, $\bar{x}$ is the global mean, and w_{ij} denotes the spatial weight between locations i and j in the spatial weights matrix. Moran's I typically falls within $[-1,1]$. Values of $I > 0$ indicate positive spatial autocorrelation (clustering of similar values), $I < 0$ indicates negative spatial autocorrelation (spatial juxtaposition of dissimilar values), and values close to 0 suggest a pattern closer to randomness. The core idea of General G is to aggregate the degree of "high-value co-occurrence" within neighborhoods at the global level, and it can therefore serve as complementary evidence to Moran's I when assessing overall spatial structure.

A Monte Carlo permutation test was used for inference, rather than relying solely on analytical methods. The spatial weights matrix stayed the same, while the observed values were randomly shuffled 999 times to create an empirical null distribution under spatial randomness. This distribution was used to calculate the simulated standardized statistic (z) and the empirical p-value (p). Permutation-based inference was chosen because streetscape sampling often has irregular spacing, local density differences, and skewed variables. In these cases, analytical tests can be sensitive to assumptions about the data and the weights matrix. Permutation tests are usually more robust since they make fewer assumptions. The results include the observed Moran's I and General G, along with z and p values; $p < 0.05$ indicates statistically significant global spatial structure.

The spatial weights matrix was created using k-nearest neighbors (KNN) weights. Unlike fixed distance-band weights, KNN gives each observation the same number of neighbors. This avoids imbalance caused by uneven point density, which often occurs along road networks with both crowded and sparse areas. A baseline of $k = 8$ was chosen to balance local detail and statistical stability. If k is too small, the weights graph can break apart, and the statistics may become unstable. If k is too large, distant neighbors are included, and local patterns get weaker. To prevent disconnection near boundaries or in sparse areas, k could be increased as needed, up to a maximum of 15, to improve connectivity while keeping neighborhoods fairly local.

To reduce artifacts from directional KNN links, the weights were made symmetric, making the neighborhood structure more like an undirected graph. The weights were then row-standardized to remove differences in total neighbor influence across observations. The dataset was also checked for problems that could affect KNN structures, such as missing values and duplicate coordinates. Missing values were removed for each variable before calculations, and duplicate geometries were dropped.

This ensures the weight matrix matches the observation vector and maintains valid neighborhood relationships.

3.6.2 Ordinary Least Squares and Geographically Weighted Regression Models

In regression analysis, Ordinary Least Squares (OLS) typically assumes spatial stationarity. In other words, it treats the relationship between predictors and the outcome as constant across the study area, with coefficients that do not change by location. For streetscape perception and color characteristics, this assumption is often difficult to justify because these variables commonly exhibit spatial dependence and heterogeneity. Street visual conditions—and how they shape emotional impressions—can differ across neighborhoods, so the strength and even the direction of relationships may vary from place to place.

Geographically Weighted Regression (GWR) is designed for this situation. It brings location into the model and estimates relationships at a local scale. Nearby observations receive higher weights, which allows coefficients to vary smoothly over space. This allows capturing spatial non-stationarity and mapping where relationships are stronger, weaker, or of different sign. OLS remains a useful global benchmark. It summarizes average effects across the full study area and provides a reference point for judging whether spatial heterogeneity is substantial.

Based on this logic, both OLS and GWR are used to examine how streetscape color complexity (C_C) and color harmony (C_H) relate to emotional perception. The dependent variables are four perception scores—beautiful, safe, boring, and depressing—and the explanatory variables are the C_C and C_H indices. Comparing OLS and GWR results allows assessment of both the global average association and the extent to which the association varies across space.

(1) Ordinary Least Squares (OLS)

The OLS regression model can be expressed as:

$$Y_i = \beta_0 + \sum_k \beta_k X_{ik} + \varepsilon_i$$

where Y_i denotes the emotional perception score of the i-th streetscape sample, X_{ik} represents the k-th color-related explanatory variable (color complexity or color harmony) for sample i, β_0 is the intercept term, β_k are the regression coefficients associated with the explanatory variables, and ε_i is the random error term.

(2) Geographically Weighted Regression (GWR)

To further capture spatial heterogeneity in the relationships between color indicators and emotional perception, this study introduces the GWR model as an extension of OLS. The general form of GWR is given by:

$$Y_i = \beta_0(u_i,v_i) + \sum_k \beta_k (u_i,v_i) X_{ik} + \varepsilon_i$$

where (u_i,v_i) denote the spatial coordinates of the i-th observation, and $\beta_k(u_i,v_i)$ are location-specific regression coefficients that vary over space. These coefficients are estimated based on a spatially weighted neighborhood centered on each observation.

During GWR implementation, all explanatory variables are standardized using z-score normalization prior to model fitting, in order to reduce the influence of multicollinearity and scale differences on the stability of local parameter estimates. In addition, to avoid singularity issues caused by duplicate observations at identical spatial locations, samples with exactly overlapping coordinates are aggregated, retaining only their mean values for subsequent analysis.

3.7 Hierarchical Multiple Regression

To test whether street color retains independent explanatory power after semantic composition is controlled, this study further applies a hierarchical multiple regression design. While the preceding OLS and GWR models examine the overall associations between color indicators and perception scores, this additional step is intended to evaluate the net effects of color beyond the scene's main structural makeup. The general form of the multiple regression model is:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i$$

where Y_i denotes the pano-level perception score for observation i, X_{ik} denotes the explanatory variables, β_0 is the intercept, β_k represents the regression coefficients, and ε_i is the error term. In this study, the dependent variables are the four pano-level perception measures—Beautiful, Safe, Boring, and Depressing—and the explanatory variables include both the whole-scene color metrics, namely color complexity (C_C) and color harmony (C_H), and the semantic control candidates derived from the segmentation results, including road, sidewalk, building, wall, vegetation, and sky.

For each perception outcome, a baseline model (M0) is first established using only semantic controls. The final semantic control set is selected separately via AIC-based forward selection, starting from an empty semantic model and adding the variable that yields the greatest improvement in model fit at each step until no further meaningful reduction in AIC is achieved. Based on M0, three nested models are then constructed: M1 = M0 + C_C, M2 = M0 + C_H, and M3 = M0 + C_C + C_H. Model performance is compared using adjusted R^2, AIC, and BIC to determine whether adding color variables improves the explanatory fit beyond the semantic baseline. For the joint model M3, a nested F-test is further used to assess whether the color block as a whole provides a significant improvement over M0. To reduce the influence of possible heteroskedasticity, coefficient inference is based on HC3 robust standard errors. This hierarchical regression framework, therefore, serves as the primary method for testing whether streetscape color adds an explanatory layer to street-level emotional perception beyond semantic composition.

4 Results and Analysis

4.1 Comparing Whole-Scene and Structure-Background Color Metrics

To assess whether color metrics are consistent across spatial-semantic scopes—and to measure both the direction and size of any differences—this study compares color complexity (C_C) and color harmony (C_H) computed under the whole-scene and structure-background scopes. The structure-background scope computes metrics only within the structure-background area defined by the semantic mask, so it focuses on the combined color characteristics of structural streetscape elements. The whole-scene scope, by contrast, summarizes the composite color structure of the entire streetscape view.

4.1.1 Analytical framework and evaluation metrics

Because "high correlation" does not necessarily imply "numerical agreement," we adopt a three-step complementary framework to assess both differences and agreement between the two scopes:

A) **Difference distribution**

For each sample, we compute the difference:

$$diff = struct - whole$$

and use a histogram to characterize the central tendency, skewness, and tail range of the differences (Table 1).

Table 1. Difference Distribution of Color Metrics:
Whole-Scene vs. Structure-Background

Color Index	mean_diff	std_diff	median_diff	q05	q95
C_C	0.008	0.037	0.003	-0.040	0.075
C_H	-0.002	0.041	-0.010	-0.050	0.074

B) **Trend consistency**

We compute both Spearman's ρ (more suitable for non-normal or monotonic relationships) and Pearson's r (as a linear-correlation complement). We further visualize the overall correspondence using a 2D histogram (log(count)) with a 1:1 reference line. This step addresses the question: Do the two scopes yield consistent ranking and overall variation trends?

C) **Agreement via Bland-Altman**

We use the Bland-Altman approach to evaluate whether the two scopes can be treated as the "same measurement" in terms of agreement:

$$mean = \frac{whole + struct}{2},\ diff = struct - whole$$

We compute the systematic bias ($bias = \overline{diff}$), and visualize how differences vary with the mean using a hexbin plot. This step answers: Is there any systematic shift?

4.1.2 Color complexity (Cc)

(1) Direction and magnitude of differences (Figure 13)

For C_C, the mean difference is 0.008, and the median is 0.003. This means C_C_struct is usually just slightly higher than C_C_whole, and the change is small; most values are close to zero. The standard deviation is 0.037. About 90% of the samples fall between -0.040 and 0.075 (q05-q95), so most differences are limited, though a few are higher.

The structure-background method often reduces the influence of large, uniform color areas, such as the sky or distant background. This means the analysis focuses more on building facades, roads, and related background features. When the main colors in this area are more evenly distributed, the measured complexity can be a little higher.

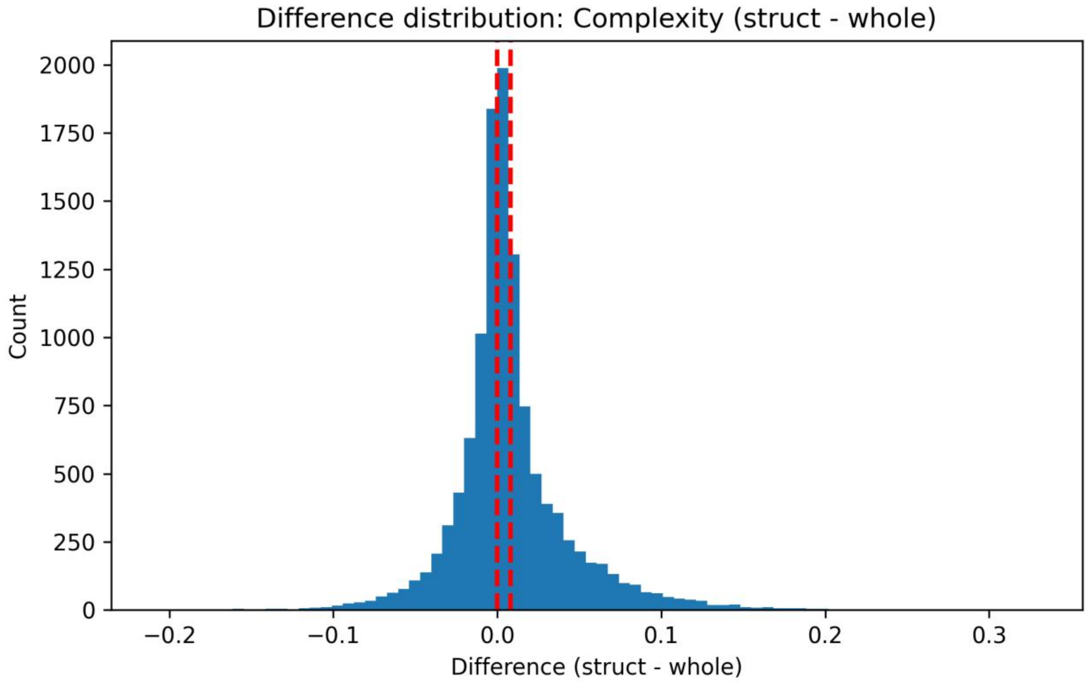


Figure 13. Difference distribution of Color Complexity: Structure-Background VS Whole-Scene

(2) Trend consistency (Figure 14)

The 2D histogram shows that the point cloud broadly aligns with the 1:1 line. The correlation coefficients are:

- *Spearman* $r = 0.787$ ($p \approx 0$)
- *Pearson* $r = 0.764$ ($p \approx 0$)

This shows strong consistency across scopes, with only moderate sample-level differences. In practice, the two are broadly similar and can be used interchangeably, as both reflect similar patterns of color complexity, though with slight differences in focus.

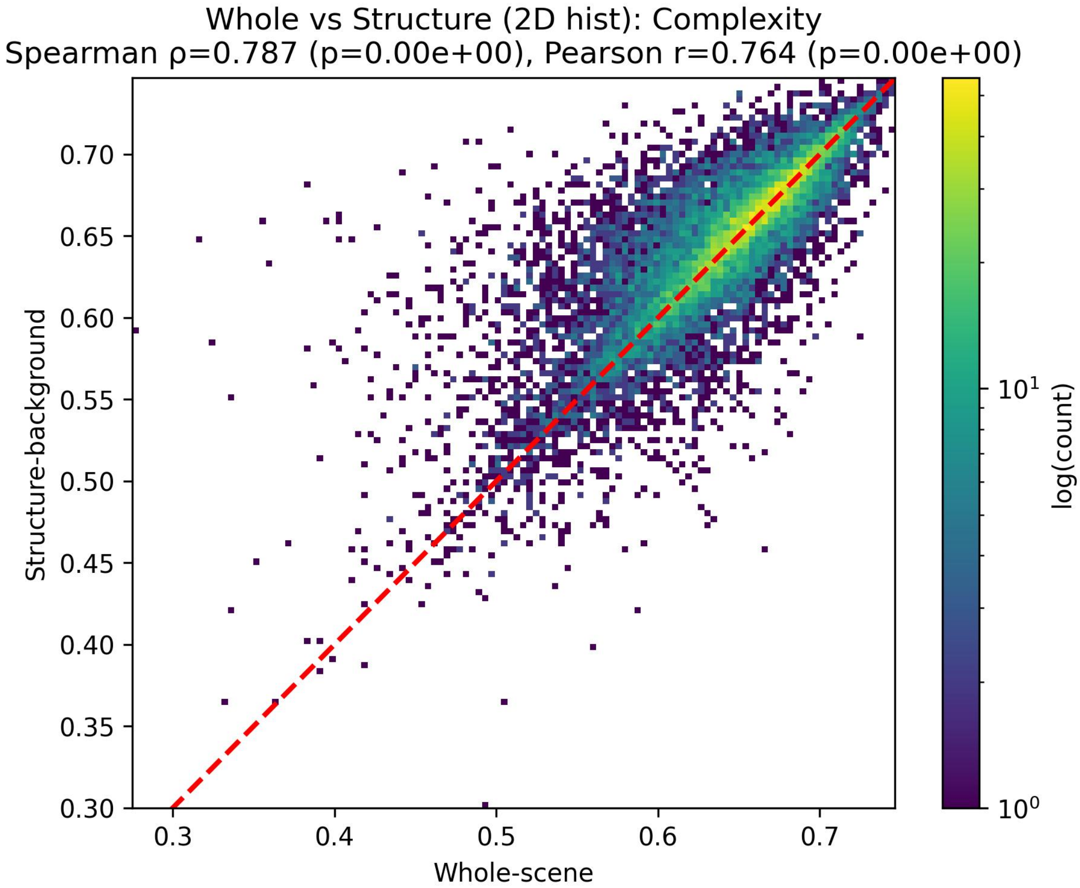


Figure 14. 2D histogram of Color Complexity (Structure-Background VS Whole-Scene)

(3) Agreement boundaries (Figure 15)

The Bland-Altman results show a systematic bias of 0.0082. The near-zero bias suggests no pronounced global systematic shift. Figure 15 also indicates a degree of heteroscedasticity: differences are more dispersed at lower mean C_C and become more concentrated at higher mean C_C. This implies that when overall streetscape complexity is low, switching scopes (whole vs. struct) affects C_C less stably; whereas in highly complex scenes, the two scopes yield more consistent C_C values.

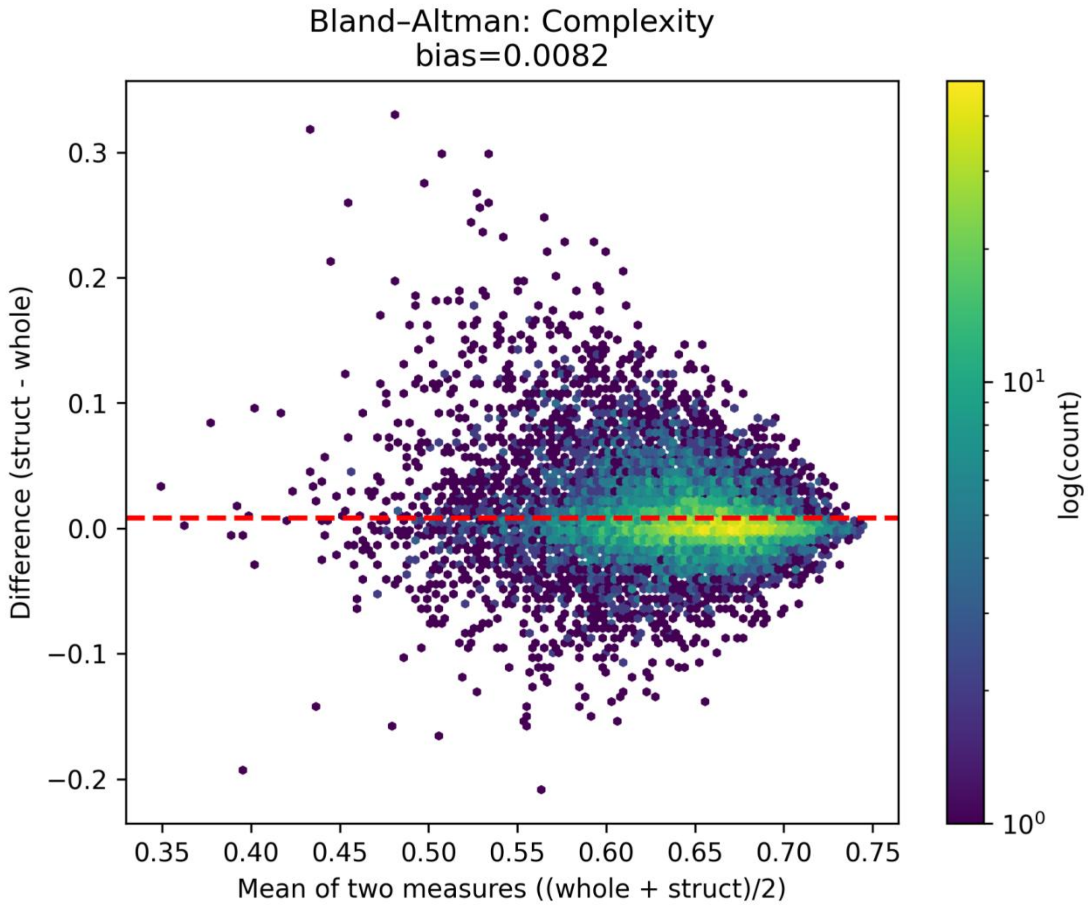


Figure 15. Bland-Altman agreement analysis of Color Complexity (Structure-Background VS Whole-Scene)

4.1.3 Color harmony (C_H)

(1) Direction and magnitude of differences (Figure 16)

For C_H, the mean difference is -0.002, and the median is -0.010, indicating that C_H_struct is slightly lower than C_H_whole with a small magnitude. The standard deviation is 0.041, and the q05-q95 range is [-0.050, 0.074]. Most samples fluctuate near zero. Compared to C_C, the C_H difference distribution is a small shift nearly centered on zero, without a systematic trend. This implies that, although the structure-background scope changes the pixel set, C_H is less sensitive to scope because it relies on a global measure of weighted inter-dominant-color differences, giving greater stability.

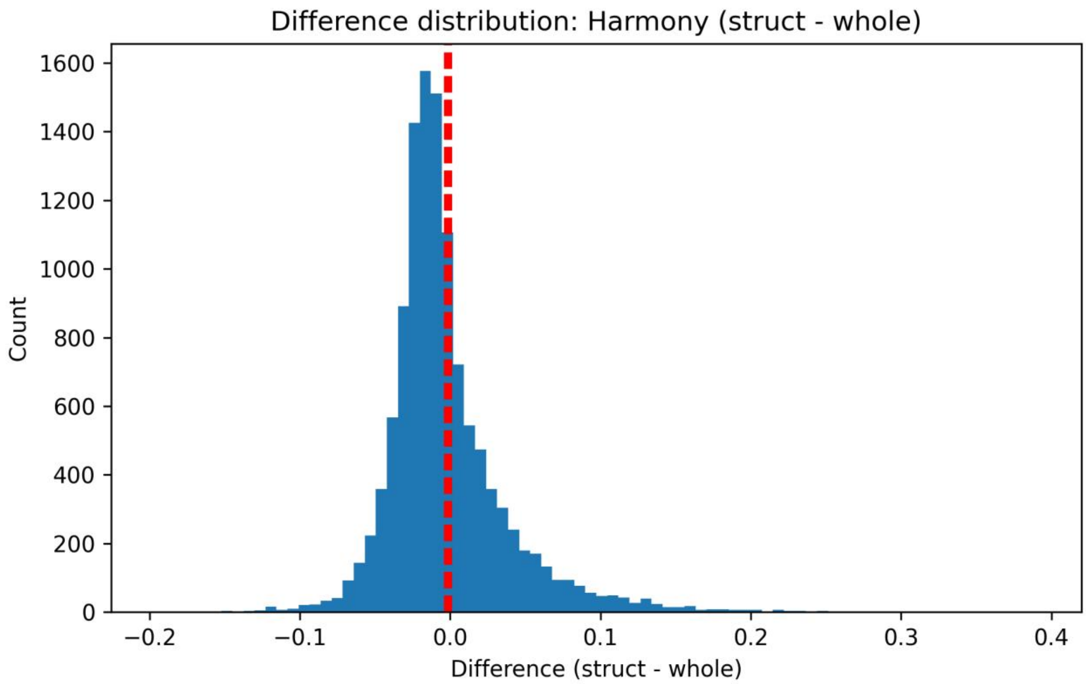


Figure 16. Difference distribution of Color Harmony: Structure-Background VS Whole-Scene

(2) Trend consistency (Figure 17)

The 2D histogram for C_H lies almost directly on the 1:1 line, with extremely high correlations:

- *Spearman* $\rho = 0.968$ ($p \approx 0$)
- *Pearson* $r = 0.968$ ($p \approx 0$)

This shows that the two scopes produce almost identical ordering and relative variation in harmony; the harmony values from structure-background and whole-scene change nearly in lockstep.

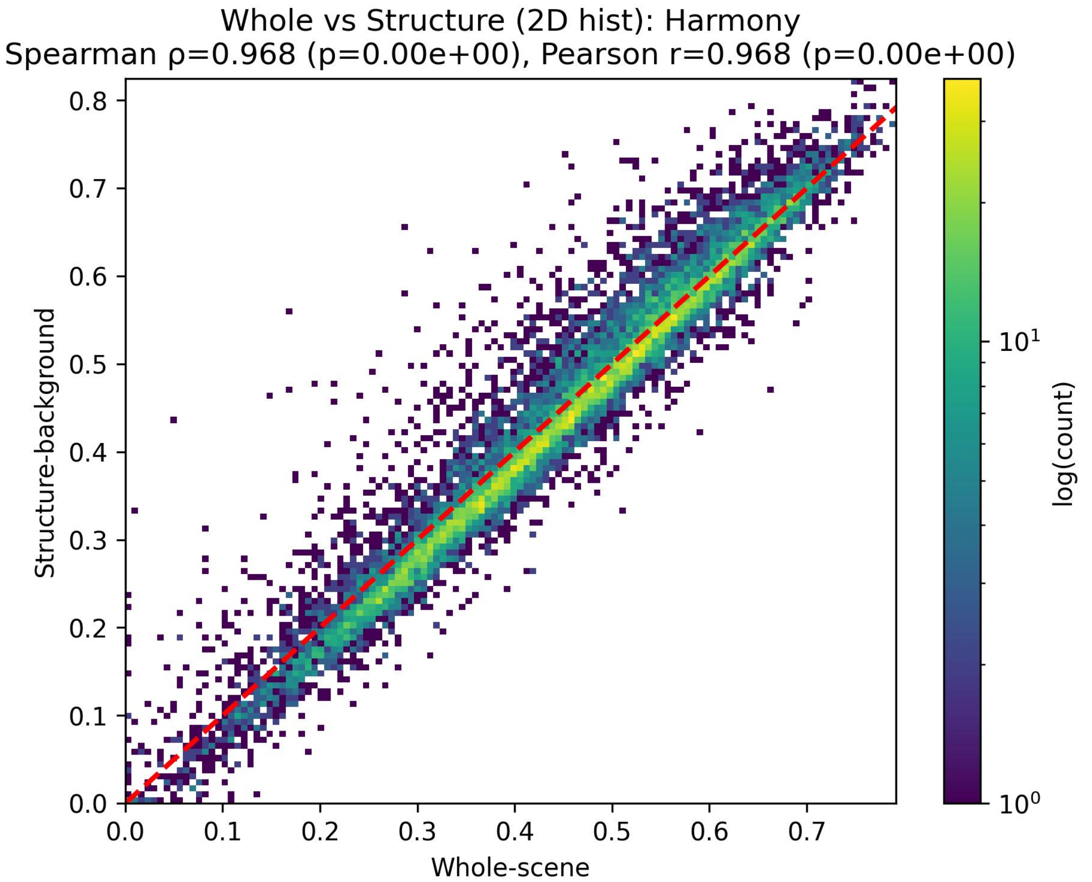


Figure 17. 2D histogram of Color Harmony (Structure-Background VS Whole-Scene)

(3) Agreement boundaries (Figure 18)

The Bland-Altman bias is -0.0024. This bias is very small, so there is no meaningful overall shift between scopes. Figure 18 also shows a mild mean-difference pattern, consistent with weak proportional bias. When C_H is low, the differences tend to be slightly more negative. As C_H increases, the differences move toward zero and may become slightly positive. This suggests that under more extreme color-combination conditions, switching the semantic scope can slightly change how dominant-color differences are weighted. Overall, the effect remains modest.

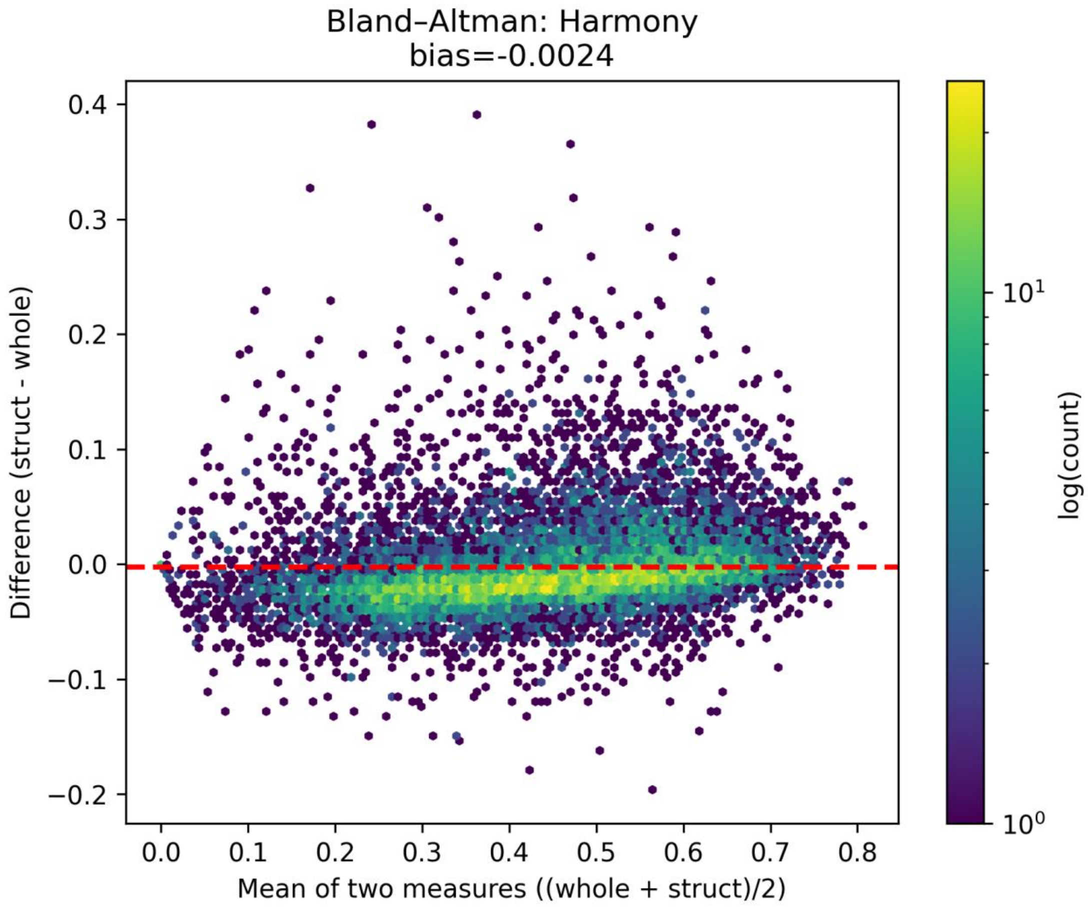


Figure 18. Bland-Altman agreement analysis of Color Harmony (Structure-Background VS Whole-Scene)

Looking at panels A-C together, the results show that whole-scene and structure-background color metrics mostly agree. The differences are small, with no strong pattern of change. Agreement for C_H is especially high (*Spearman* $\rho \approx 0.968$; *Pearson* $r \approx 0.968$). C_C has a wider range (*Spearman* $\rho \approx 0.787$), but the average difference is still small (*bias* $\approx +0.008$). Most differences are close to zero. Most samples fall within a narrow range. In short, both approaches capture the same color-structure signal, though there is some variation between samples.

Building on these findings, to maintain consistency, the following modeling and spatial analyses use the whole-scene color metrics (C_C_whole and C_H_whole) as main inputs. This approach has four supports:

1. High agreement and small bias show that the main conclusions are unlikely to change. Both rank and linear associations, as well as Bland-Altman results, confirm strong consistency. Using a single scope avoids duplication and preserves the core signal.
2. Whole-scene metrics match the measurement domain of the perception outcomes. Perception scores are predicted from the full image. If predictors are based only on the structure-background, but outcomes cover the whole scene, the model mixes scopes. Using whole-scene metrics aligns predictors and outcomes, improving interpretability.
3. Whole-scene metrics better reflect what people actually see. Emotional and aesthetic responses are based on the full field of view, not just internal structure color. Whole-scene metrics include sky, vegetation, and open background, approximating real visual exposure and aligning with common practice.
4. Using a single scope makes models simpler and more interpretable. Including both types, especially when highly correlated (as with C_H), introduces redundancy and risks issues such as multicollinearity. Whole-scene inputs keep analysis cleaner and more stable.

While structure-background metrics remain useful for specific questions—such as determining whether the organization of color within structural elements affects perception—the main goal here is to build a clear color-perception framework using the same visual scope as the perception prediction. For this reason, the next chapters use C_C_whole and C_H_whole as the main explanatory variables, the later discussion of C_C and C_H refers to these whole-scene measures.

4.2 Spatial Distribution Patterns of Street Color

Building on the previous section, which established consistent use of the whole-scene color indicators (C_C_whole and C_H_whole), this section maps those indicators onto Zurich's road network to reveal spatial heterogeneity in streetscape color and its relationship with functional urban spaces. Figures 19 and 20 show the classified results of color complexity and color harmony across road samples. Class thresholds, defined using the Natural Breaks (Jenks) method, align with the data’s clustering structure and non-uniform distribution, thereby enhancing the readability and interpretability of spatial patterns.

4.2.1 Color Complexity (Cc)

As shown in Figure 19, the spatial pattern of Color Complexity (C_C) does not fit a simple “center-periphery decrease,” but it also does not read as a clean reversal with “weak clustering in the center and dominance at the edges.” High-complexity segments (red) run more continuously through the dense street network of the city center and along several radial arterial corridors, creating chain-like bands that cut across the built-up area. At the same time, clusters of high values also appear on some peripheral segments, including roads in the southwestern part of the city and routes next to forested or sloped terrain. Taken together, the map points to a corridor-driven pattern: where traffic organization is stronger, interface types are more mixed, and street elements accumulate, the streetscape tends to contain vegetation, road and traffic infrastructure, building facades, and other components at once, often with clear edges and strong color contrasts. These combinations break the scene into more distinct color patches, which pushes C_C upward.

By contrast, low- to moderate-C_C segments (blue and yellow) appear as embedded, discontinuous patches within the network rather than forming a stable cluster in a single area. In the historic core (Altstadt), traditional materials and facade systems often favor overall coherence and visual continuity, which would be expected to limit extreme color proliferation. Yet Figure 19 shows many medium-to-high segments in and around this area, suggesting that dense street grids, active commercial frontages, and traffic nodes can still sustain a relatively large number of color sources and strong local contrasts. A similar point applies along the shores of Lake Zurich and the

Limmat River. Water surfaces and sky can introduce large, stable color fields in certain views, but the complexity of lakeside streets depends more on the combined influence of vegetation, building interfaces, road infrastructure, and activity intensity near the shoreline. As a result, the map shows a mix of low, medium, and high values rather than a clear trend toward "overall lower complexity along the lakeshore."

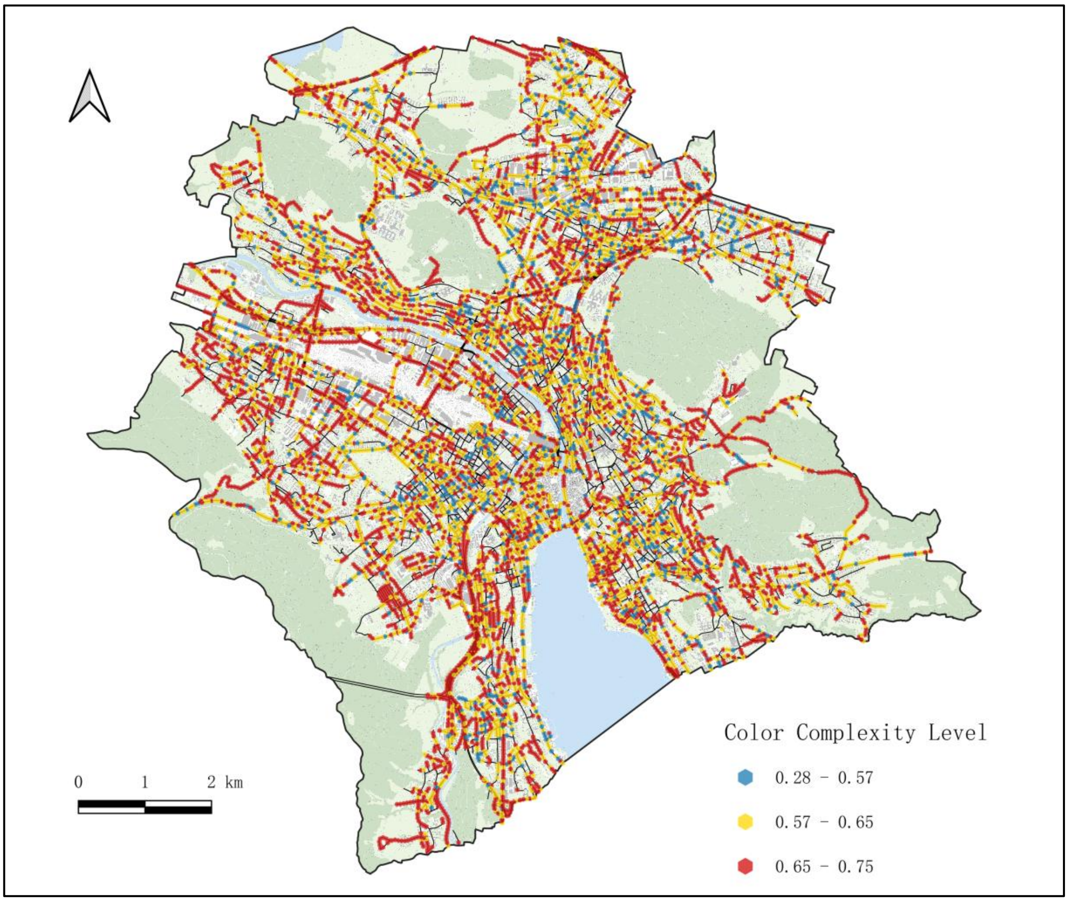


Figure 19. Visualization of color complexity

4.2.2 Color Harmony (C_H)

Figure 20 shows that Color Harmony (C_H) values are generally higher in the city center and fluctuate more toward the edges, but the pattern does not form a perfect circle. High-harmony segments, shown in red, are more continuous in the core urban road network and stretch outward along major arterial roads. In contrast, roads on the outskirts show a mix of high, medium, and low values. This may be because the historic center has a more stable set of materials and facade designs, along with stricter design controls for public spaces. These factors help reduce clashes of color and allow central streets to maintain longer stretches of medium-to-high harmony.

Lower harmony segments, shown in blue, are more common along certain edge corridors and nodes, where they appear as scattered patches. These areas often combine different types of interfaces and infrastructure, such as various commercial signs, transport systems, and modern facade materials. This mix makes it harder for nearby elements to maintain a consistent color scheme, increasing the risk of color clashes. Waterfront roads and streets near green spaces do not always have high harmony either; instead, they show a mix of high, medium, and low levels. Sometimes, water or greenery provides a broad background color that helps smooth transitions. Still, harmony along waterfront streets largely depends on the nearby built environment, the road's significance, and the amount of shoreline infrastructure.

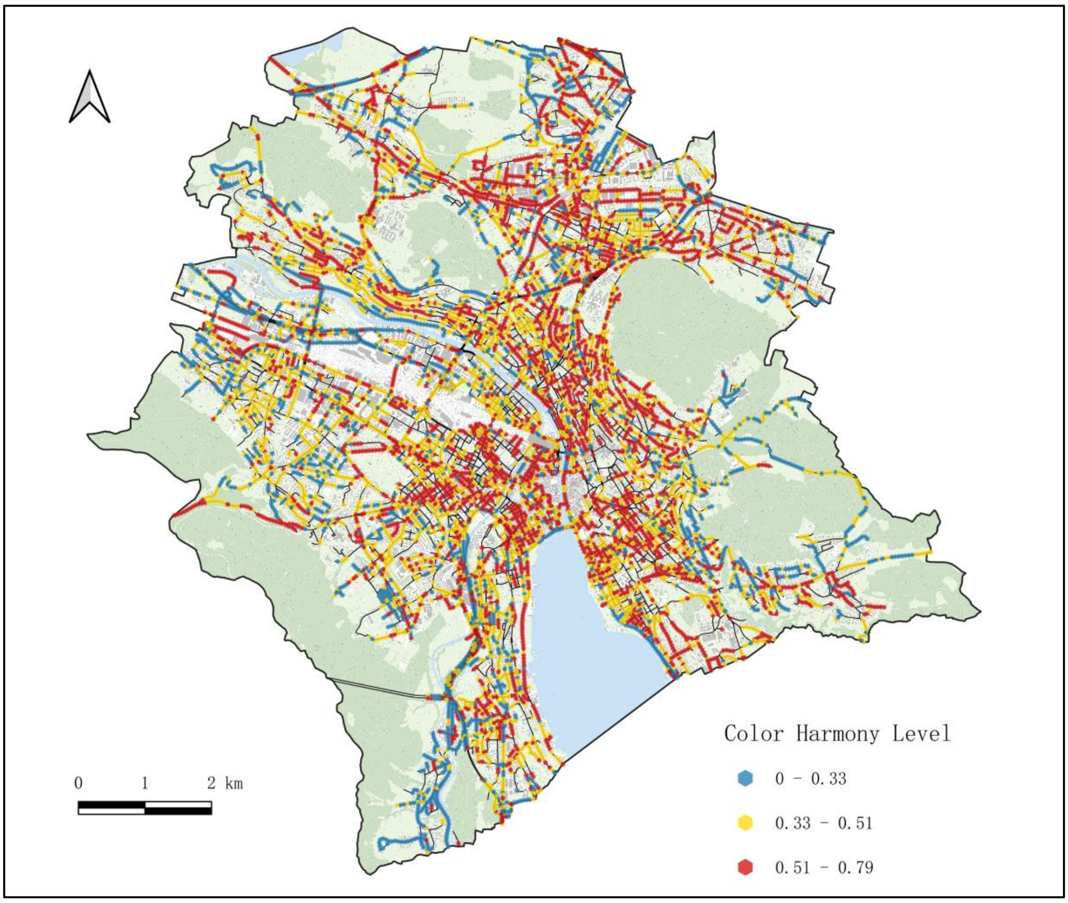


Figure 20. Visualization of color harmony

4.3 Emotion Perception Analysis

4.3.1 Emotion Correlation Analysis

To describe how the emotion-perception dimensions relate to each other, a Pearson correlation analysis was run on four scores—beautiful, safe, boring, and depressing—and the results are shown in a correlation bubble plot (Figure 21). In the figure, bubble color marks the direction of association (red = positive, blue = negative), and bubble size indicates its strength (larger bubbles = stronger correlations).

Overall, the two “positive perception” indicators show strong consistency: beauty and safety exhibit a relatively strong positive correlation (0.67), indicating that, in the streetscape samples, locations rated as more beautiful are also more likely to be rated as safer. In contrast, there is also a degree of synchronous variation between “negative perception” indicators: boring and depressing show a moderate positive correlation (0.31), suggesting that more boring environments are often accompanied by a stronger depressive experience. Between positive and negative dimensions, the correlations are clearer and stronger. Depressing shows a relatively strong negative correlation with beautiful (-0.61) and also a relatively strong negative correlation with safe (-0.63), implying that when an environment is perceived as more depressing, its beauty and safety ratings are usually lower. In addition, boring has a moderate negative correlation with safe (-0.42), but only a weak negative correlation with beautiful (-0.18). This suggests that the opposition between “boring” and “safe” is more pronounced in the sample, whereas its linear association with “beautiful” is relatively weak, potentially reflecting more complex nonlinearity or context dependence.

Furthermore, if beautiful and safe are categorized as relatively positive emotion perceptions, and boring and depressing as relatively negative emotion perceptions, one can observe that correlations are positive within the positive group, positive within the negative group, and generally negative between the positive and negative groups. This structural pattern indicates that residents' emotional evaluations of streetscapes are not independent of one another, but instead display a certain "positive-negative" dimensional differentiation and coordinated variation.

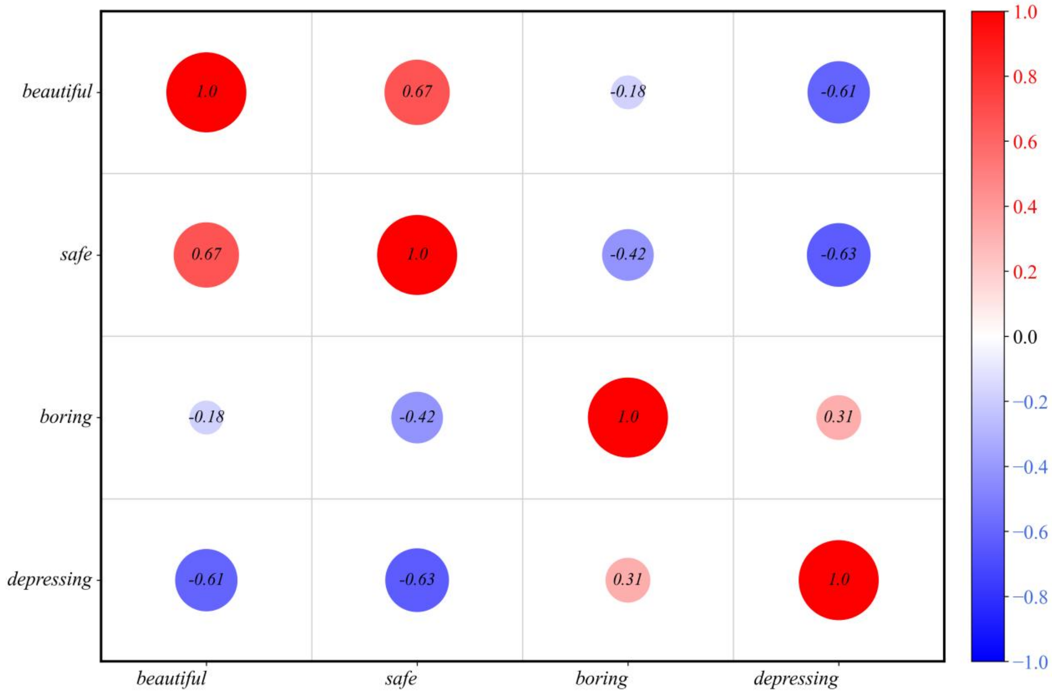


Figure 21. Bubble correlation matrix of the four perception scores

4.3.2. Spatial Distribution Patterns of Emotional Perception

To further understand the intensity of residents' perceptions across the four emotional dimensions, this study applies the Natural Breaks (Jenks) classification method to divide each emotional perception score into three levels. Four distinct colors are used to differentiate the four types of emotional perception (see Figures 22-25).

In Zurich, medium-to-high Beautiful values appear most often on streets where public life is concentrated and where views open to greater landscape exposure. This is especially clear on several major and secondary arterials in the central area, and on the corridors that continue outward from the center, where higher values appear in longer, more connected stretches. Some waterfront streets also connect to these corridors, forming fairly steady “good” routes rather than scattered, one-off segments. That pattern makes sense given how people use the city: walking and leisure in Zurich are closely tied to the shores of Lake Zurich and the Limmat River. The lakeside promenade and the Limmat riverfront are major public spaces, and when these waterfront settings meet the high-quality street edges of the historic core, the overall experience tends to feel more attractive—so higher beauty scores are more common. Lower Beautiful values, in comparison, appear more often on segments built around traffic movement, where the scale is larger, and the street interface is more uniform. When what dominates the scene is road space, traffic facilities, and other infrastructure, the visual experience becomes more repetitive, and it is harder for streets to consistently receive high beauty ratings.

On the Safe map, the medium-to-high segments span a fairly large area. They become more continuous and denser in the inner city, the northern lakeside area, and several major activity corridors, and they broadly line up with the pattern seen for Beautiful. One likely reason is that streets with more complete interfaces, clearer spatial order, and steadier public activity are read by the model as safer. Lower values, in contrast, occur more often along a set of transport corridors in the west-northwest, along northern peripheral roads, and along a north–south corridor in the southwest. In these areas, they appear as linear runs or broken pieces sitting within an otherwise high-value background.

The Boring map, by comparison, shows a clearer “lower in the center, higher toward the periphery” pattern. Lower boredom scores cluster in the inner city and nearby neighborhoods where accessibility is high, and street activity is dense. Higher Boring values, in contrast, tend to run for longer distances on peripheral roads and on corridors that serve transport more strongly and sit in areas with relatively single land-use structures. This split fits differences in how much information and activity streets carry. Where commercial displays are more present, public transport access is better, and pedestrian activity is stronger, the streetscape offers more cues and more variation, so boredom is less likely to dominate. Where roads pass for long stretches through settings with a single function and little change at the street edge, higher Boring values are more likely to extend and link up along the network.

The spatial distribution of depressing looks less like a simple gradient and more like a mix of patchy clusters and corridor-like bands. Compared with Boring, higher depression scores do not spread as large, uniform areas. Instead, they appear as broken, high-value patches embedded in the road network. These patches are not random: they often line up along a few corridors with higher traffic capacity and stronger links between districts, which makes the higher values traceable as linear bands. This pattern points to a different setting for Depressing—streets where spatial cues feel less steady and where the environment changes more often, rather than a general increase concentrated in either the city center or the periphery. Zurich-West offers a concrete example. Official descriptions present it as a diverse district reshaped from former industrial areas, where industrial heritage sits next to strong architectural contrasts and an active mix of cultural and entertainment functions. A place built around sharp contrasts and quick transitions provides one plausible lens for why some Depressing segments rise locally in this area.

Taken as a whole, the four emotion patterns suggest that street-level emotional perceptions in Zurich are not randomly distributed. Instead, they show a fairly stable typology: higher positive scores (Beautiful, Safe) are more common in the inner city, along waterfront areas, and on streets with higher-quality urban interfaces, while higher negative scores (Boring, Depressing) show up more often at the urban edge, on transport corridors with relatively single functions, and on linear road segments where spatial transitions are more frequent.

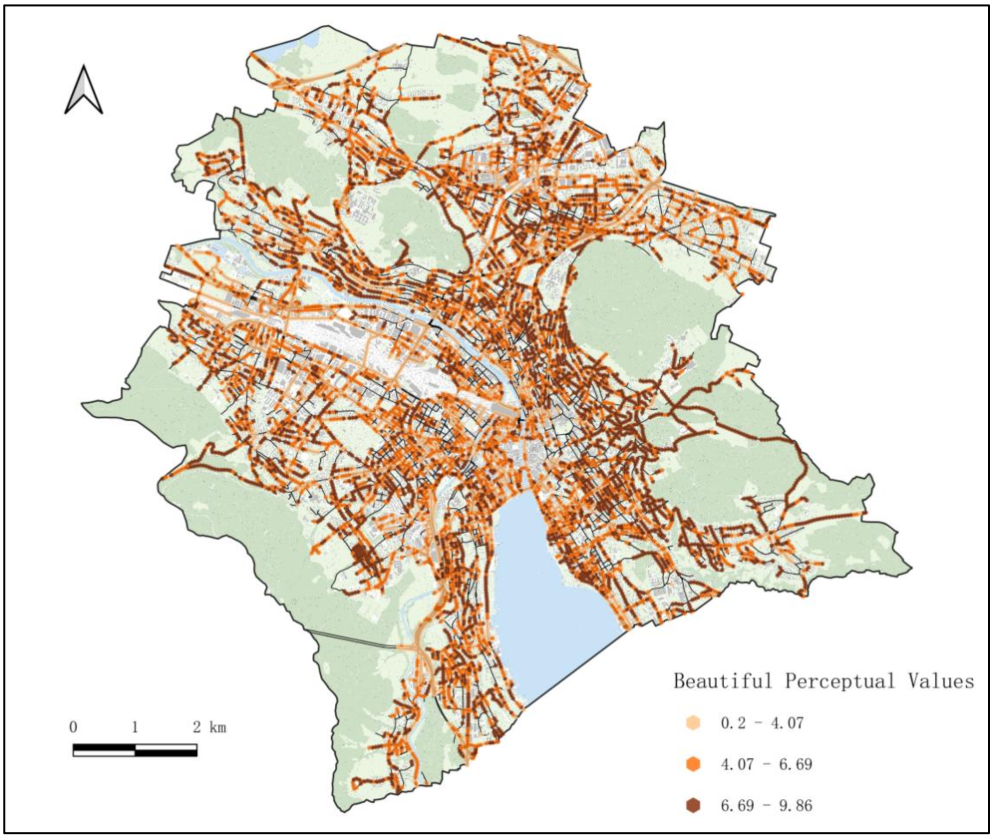


Figure 22. Visualization of Beautiful perception

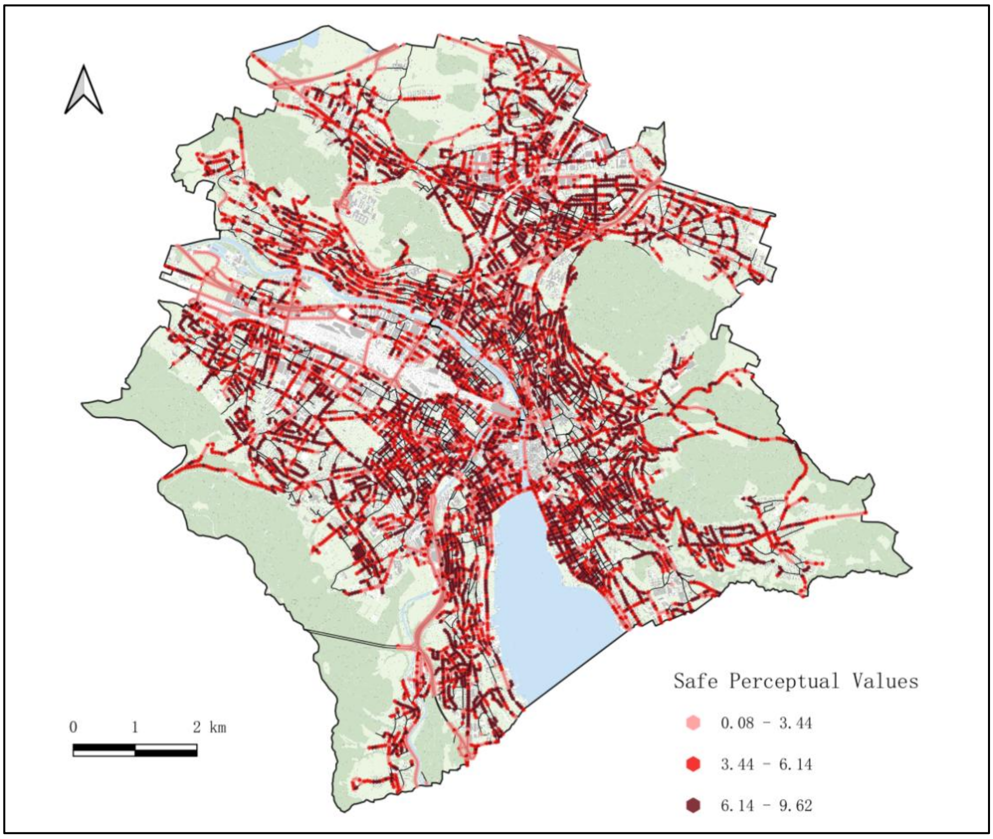


Figure 23. Visualization of Safe perception

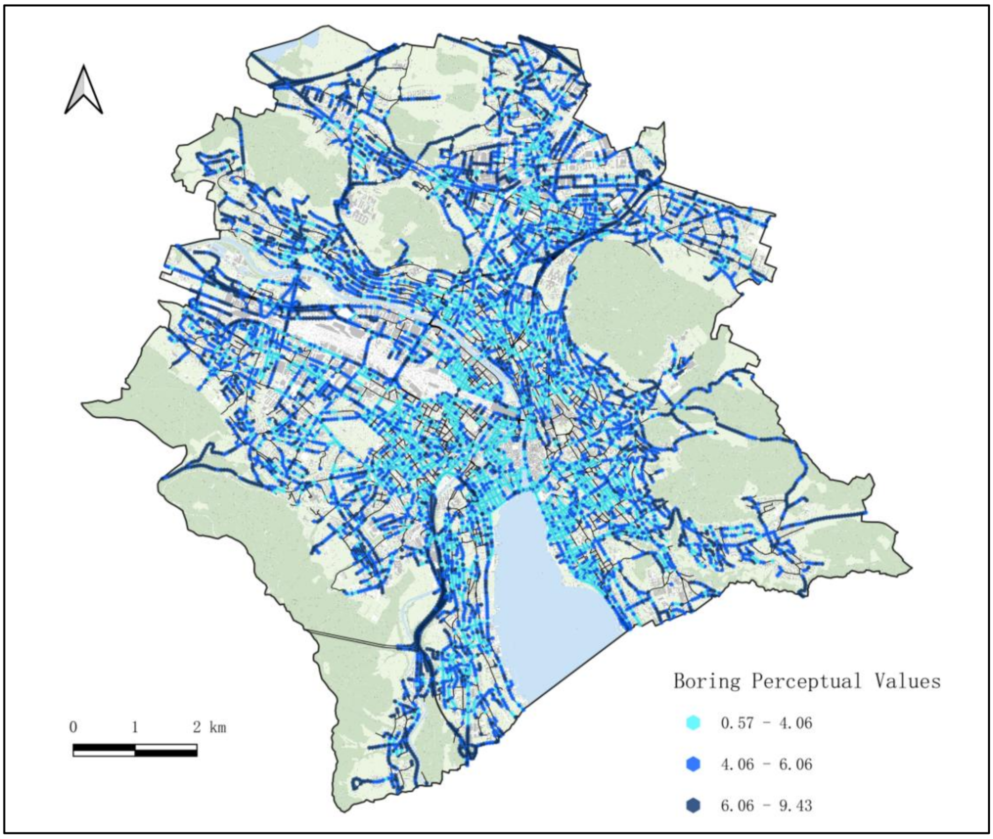


Figure 24. Visualization of Boring perception

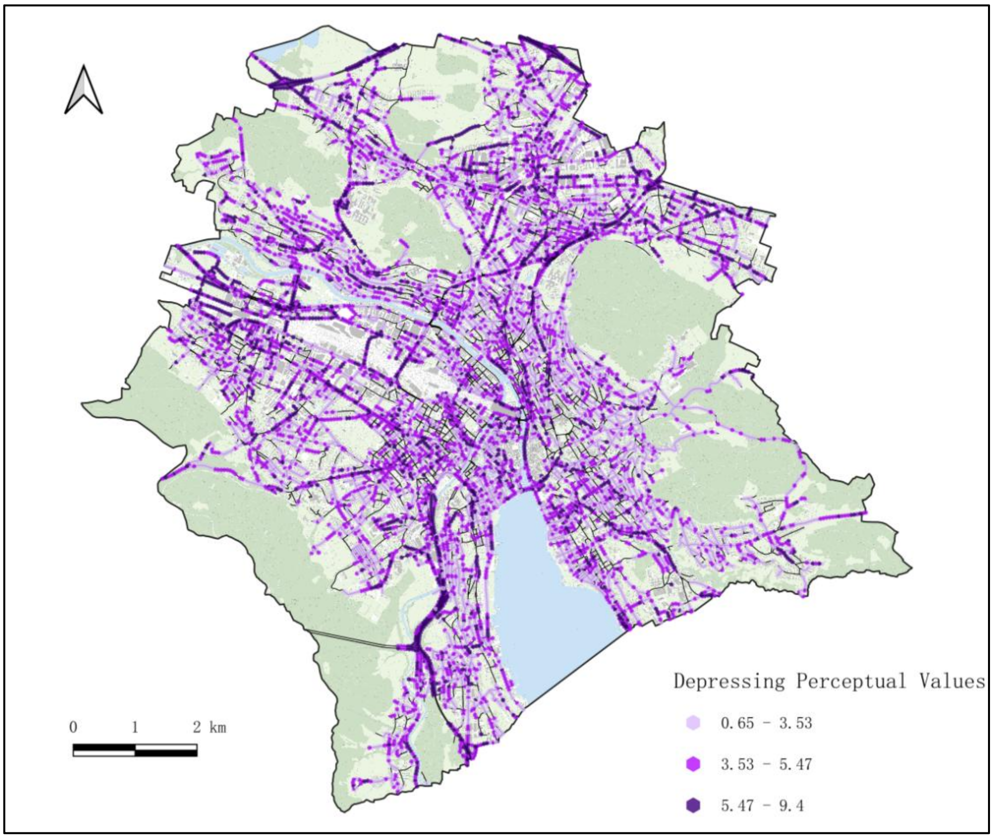


Figure 25. Visualization of Depressing perception

4.3.3 Spatial Clustering Patterns of Emotional Perception

Because spatial elements exhibit proximity effects and mutual interaction, street samples that are closer in space are more likely to show similar levels of emotional perception. To test the clustering patterns of emotional perceptions across Zurich's road network, this study constructed a k-nearest-neighbor (KNN) spatial weights matrix based on point data and, on this basis, computed the global Moran's I to evaluate the strength and significance of spatial autocorrelation in four emotional perception scores—Beautiful, Safe, Boring, and Depressing (Table 2). In building the weights, we set $k = 8$ as the baseline and adopted an adaptive strategy that increases k to ensure the connectivity of the weights network; ultimately, all four emotion variables achieved connectivity at $k = 9$, and the sample size used for computation remained consistent after removing missing values and duplicate coordinates. The results show that Moran's I for all four emotional variables is significantly positive, with permutation-test p values all equal to 0.001, indicating that their spatial distributions deviate significantly from randomness and exhibit stable clustering structures. Among them, Safe ($I = 0.478$) and Beautiful ($I = 0.455$) show the strongest spatial autocorrelation, suggesting that "positive perceptions" are more likely to form continuous high-/low-value patches along the road network; Boring ($I = 0.416$) and Depressing ($I = 0.327$) also display significant clustering, but with relatively weaker intensity, implying that clustering in "negative perceptions" exists yet is more susceptible to local scene-level variation.

It should be noted that the global Moran's I can only reveal overall "like-with-like" clustering and cannot distinguish whether such clustering is primarily driven by "high values" or "low values." To further identify which type of clustering the emotional perceptions tend toward at the global scale, we additionally computed the Getis-Ord General G statistic under the same KNN weighting framework and obtained z and p via permutation tests (Table 2). The results show that General G for all four emotional variables passes the significance test ($p = 0.001$), and all z values are positive, indicating that their global clustering is better characterized by a high-high clustering tendency—i.e., “high values are more likely to be near one another”—rather than being dominated by “low-value patches.” In other words, safe and beautiful not only exhibit significant positive autocorrelation but also more clearly highlight the spatial concentration of high-value areas; although boring and depressing show comparatively weaker clustering intensity in Moran’s I, the significantly positive *z* values of General G indicate that their high values likewise display a globally meaningful tendency to concentrate—namely, “high-boring/high-depressing” hotspot-like patches are more likely to emerge along the urban road network, rather than being completely dispersed or driven only by low-value clustering.

Table 2. Spatial autocorrelation and high/low clustering analysis

Emotional	**Spatial Autocorrelation Analysis**			**High/Low Clustering Analysis**		
Perception	**Moran’s I**	**Z**	**P**	**General G**	**Z**	**P**
Beautiful	0.455	115.068	0.001	0.000944	49.517	0.001
Safe	0.478	122.100	0.001	0.000967	60.014	0.001
Boring	0.416	111.932	0.001	0.000930	47.423	0.001
Depressing	0.327	83.765	0.001	0.000938	46.360	0.001

4.3.4 Hotspot Analysis

To locate statistically significant “hotspots” and “coldspots” for each emotion dimension, Getis–Ord Gi* hotspot analyses were run for the four scores—safe, beautiful, boring, and depressing—using the same KNN-based spatial weights framework described above. Hotspots and coldspots were identified at the 90% / 95% / 99% confidence levels (Figures 26-29).

In the Beautiful hotspot results, significant hotspots appear as continuous chains along the eastern and southeastern city edges, especially on roads next to forested areas and landscape boundaries. The city center and some waterfront areas also have hotspots, but these are more scattered. By contrast, cold spots are concentrated along several transport corridors in the west and extend into some northern peripheral road segments. The result is clear spatial heterogeneity: higher-value clusters form along edge roads with strong landscape features, while lower-value clusters are tied to corridors shaped by traffic. For Safe, cold spots are more continuous than hotspots. They form band-like patterns along several corridors in the west-northwest, a north–south corridor in the southwest, and parts of the north periphery. Hotspots are less cohesive and appear as multiple clustered patches in the central, eastern, and northeastern areas, with some near the lake. Overall, Safe does not reduce to a simple center-periphery split: lower values cluster along corridors, while higher values form clusters within residential and mixed-use areas.

In the Boring hotspot results, significant cold spots are widespread in the central city and nearby districts. Lower values form continuous clusters in the central–southern area and northern lakeside. In contrast, hotspots appear in chain-like patterns along the city's outer edges and on several peripheral roads with strong transport functions. Of the four dimensions, Boring shows the clearest pattern: lower in the center, higher toward the periphery. This suggests that boredom clusters in edge settings, where streets are designed for circulation and offer fewer spatial cues. The central city, with more mixed functions and richer street experiences, shows fewer strong boredom clusters. For Depressing, significant hotspots form corridors: along the northern boundary, several east-west corridors in the northwest, and a north-south corridor in the southwest. Cold spots do not form a continuous band along the lake or river, but cluster in the southeast and some central-eastern districts; noticeable pockets of lower values also appear near parts of the lakeshore. Overall, Depressing is best described as higher values along transport corridors, lower values concentrated on the southeastern side, and some open districts. This points to depressing perceptions clustering in linear road environments dominated by traffic, while appearing weaker in waterfront or more residential areas.

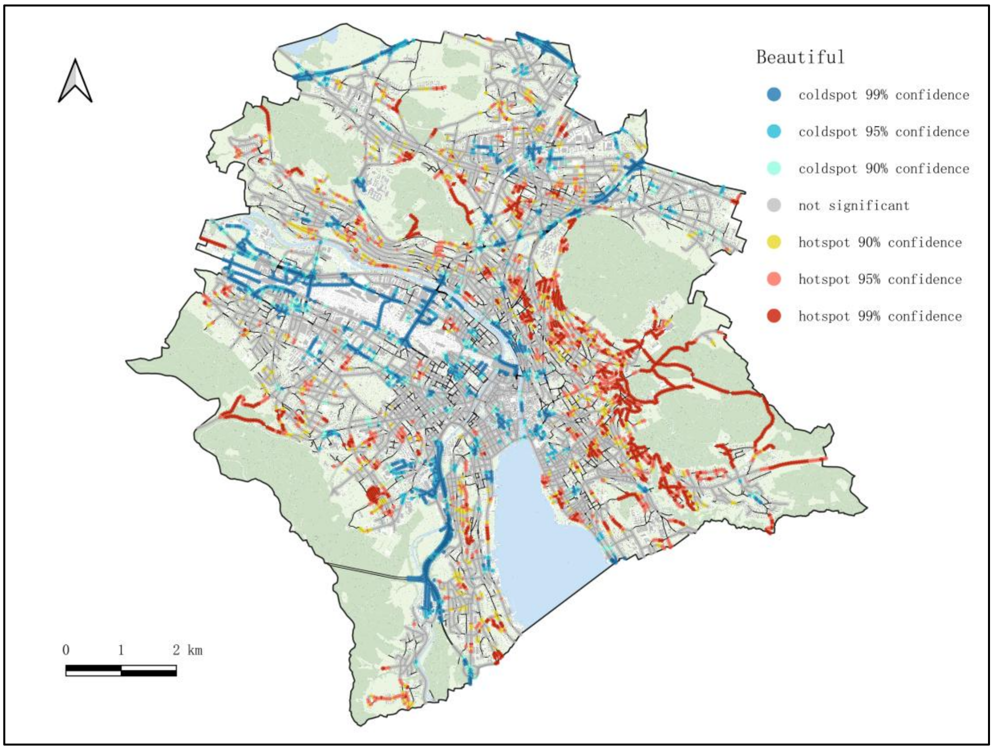


Figure 26. Distribution of Beautiful perception hotspots and coldspots

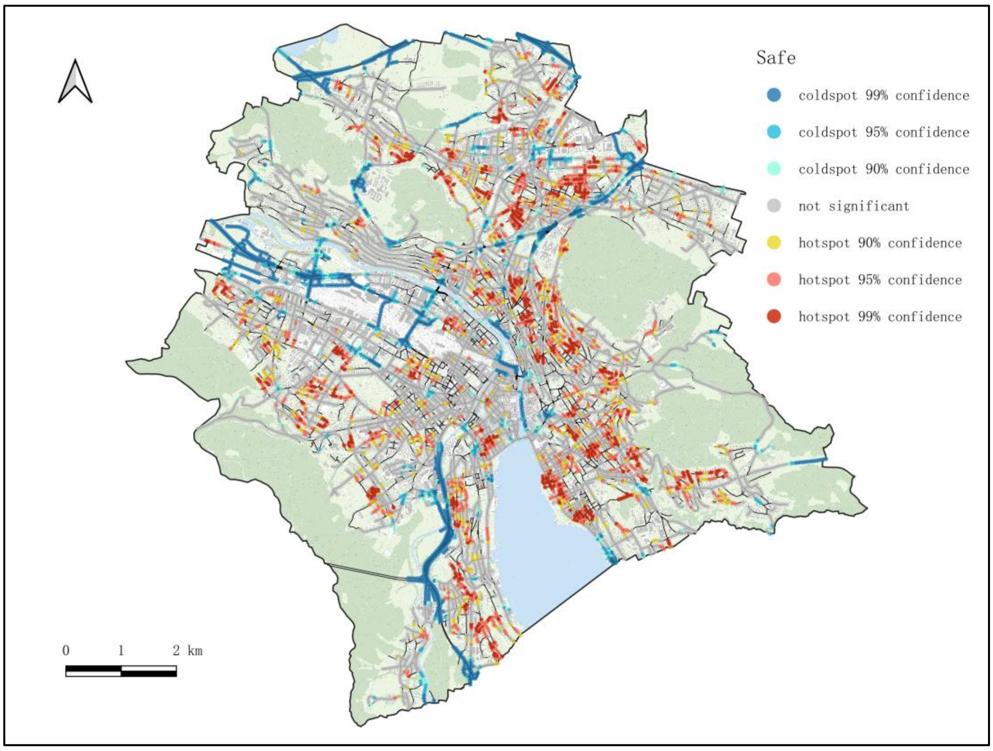


Figure 27. Distribution of Safe perception hotspots and coldspots

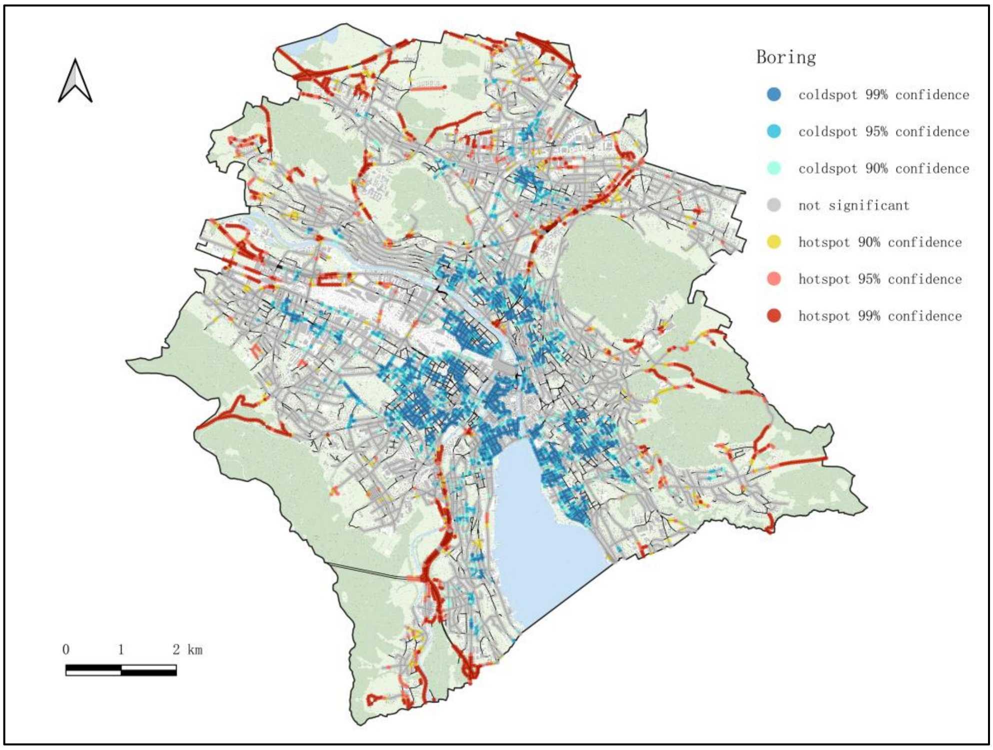


Figure 28. Distribution of Safe perception hotspots and coldspots

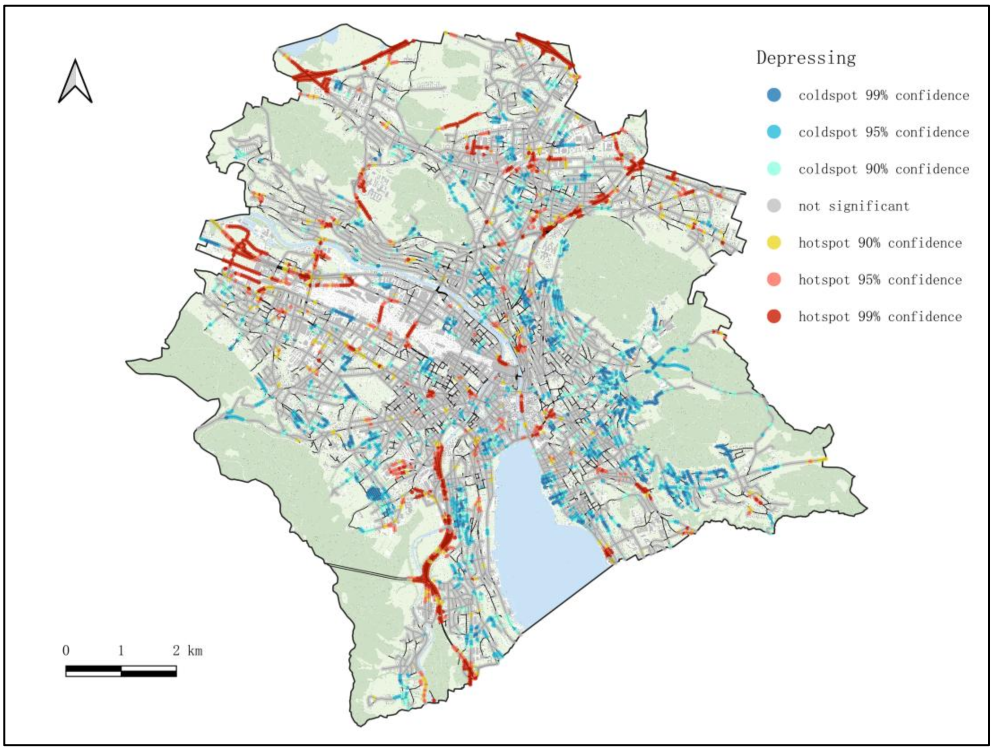


Figure 29. Distribution of Safe perception hotspots and coldspots

4.4 Analysis of Color Metrics and Emotional Perception

4.4.1 Effects of Color Complexity and Harmony on Emotional Perception

To test linear associations between streetscape color and emotional perception, Pearson correlations were computed at the pano_level between color complexity (C_C), color harmony (C_H), and four perception scores. Missing values were handled using pairwise deletion, and significance stars are based on two-tailed tests (**: $p < 0.01$; *: $p < 0.05$). Results are reported in Table 3.

Overall, C_C shows a significant positive correlation with boring and also a significant positive correlation with beautiful, while it is significantly negatively correlated with safe and depressing. C_H shows the opposite pattern: it is significantly positively correlated with safe and depressing, and significantly negatively correlated with beautiful and boring. While these relationships are statistically significant, the correlation coefficients are generally small. This points to a pattern of effects that are modest in size but relatively consistent, which motivates the follow-up models that add multivariate controls and account for spatial structure.

Table 3. Color-emotion perception correlation analysis

	Safe	Beautiful	Boring	Depressing
Color Complexity	-0.037 **	0.074 **	0.152 **	-0.024 **
Color Harmony	0.051 **	-0.111 **	-0.165 **	0.089 **

Subsequently, to further examine the potential influence of streetscape color complexity and color harmony on emotional perceptions, this study fitted separate Random Forest Regression (RFR) models for the four perception dimensions and used feature importance to quantify the relative explanatory contribution of the two color indicators (Fig. 30). For comparability, the importance scores of the two color variables were normalized within each model so that their combined contribution for a given emotion sums to 1.

The results also point to a consistent pattern across all four outcomes: color harmony shows slightly higher relative importance than color complexity. This holds for beautiful, boring, and depressing, and the gap is largest for safe. Overall, the two indicators contribute at a similar level, but the models repeatedly suggest a small advantage for harmony. Put differently, compared with the diversity of colors and their proportions, the overall fit and coherence among dominant colors appears to explain a bit more of how streets are evaluated emotionally—especially in perceived safety.

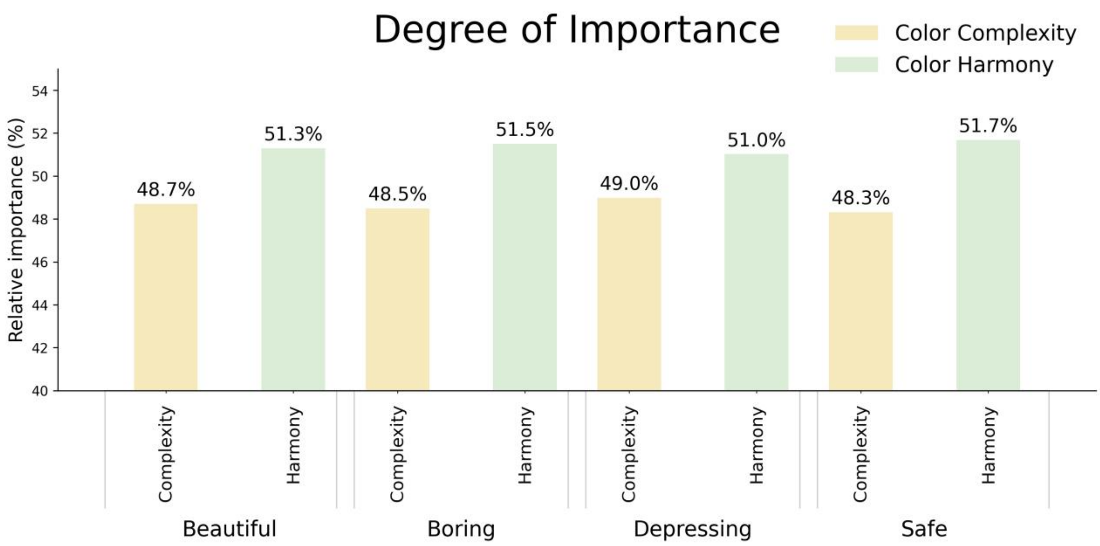


Figure 30. Importance scores of the two indicators of color for emotional impact: feature importances are model-internal relative weights and are normalized within each emotion-specific model; therefore, the importances of complexity and harmony sum to 100% for each emotion.

To better capture non-linear relationships between color metrics and emotional perception, partial dependence plots were generated from the random forest regression models (Figure 31). These plots show how the predicted emotion score changes as a single color indicator varies. They average out the influence of the other variables. Across the four outcomes, both C_C and C_H show clear non-linear patterns, including threshold-like shifts and signs of diminishing returns.

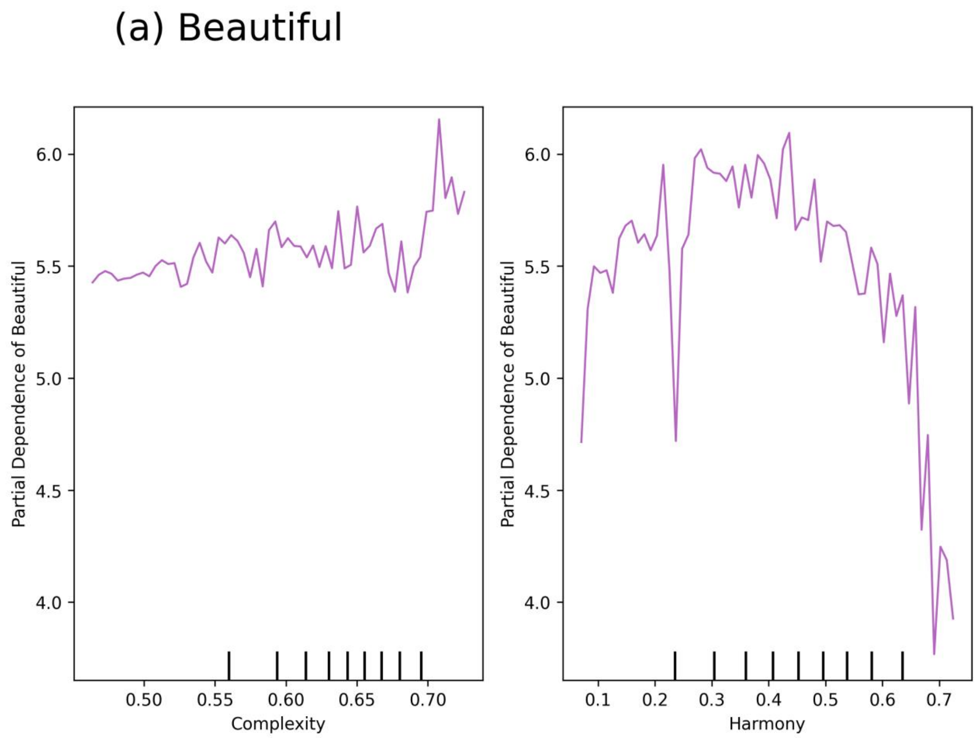
For beauty, the response to color complexity is mostly flat, with only a small increase at the highest C_C values. This suggests a limited role of complexity in predicted beauty. Very high complexity may still add a small gain. The pattern of C_H is stronger and follows an inverted U shape. Beauty increases quickly from low harmony, stays high at moderate harmony, and then drops in the upper range. This implies that moderate matching supports beauty. Very high harmony—often closer to uniformity—may align with lower beauty in the model.

For safe, C_C shows a gentle downward trend, with a clearer drop at the high-complexity end. Highly fragmented color environments are less likely to be predicted as safe. C_H follows a rise–plateau–decline pattern. Safety increases from low to moderate harmony, remains fairly stable, and then decreases again at very high harmony. Improved matching helps to a point, but the benefit weakens—and may reverse—when harmony becomes extreme.

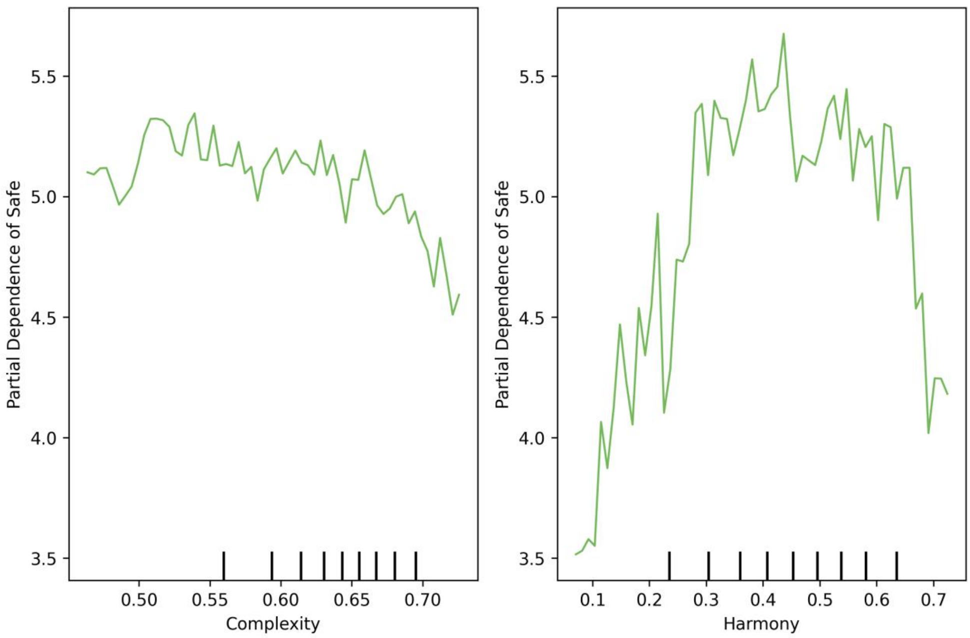
For boring stimuli, Cc increases in a largely monotonic manner. Boredom is fairly stable at low complexity, then rises more clearly as Cc increases. In the model, higher color diversity and more dispersed color proportions align with higher boredom. By contrast, C_H shows a U-shape. Boredom is highest at both low and high harmony, and lowest at moderate harmony. This suggests that both strong mismatch and strong uniformity are associated with boredom. A balanced palette corresponds with lower boredom.

For depressing, the marginal effect of color complexity is weaker overall. It drifts upward at higher Cc. This indicates that complexity is not a strong driver of depressing scores, though very high complexity can coincide with slightly higher predictions. C_H again shows a U-shaped pattern. Depressing scores are lower around moderate harmony and higher at both ends, with a stronger increase on the high-harmony side. This suggests that depressive perceptions are more sensitive when harmony moves away from moderate levels.

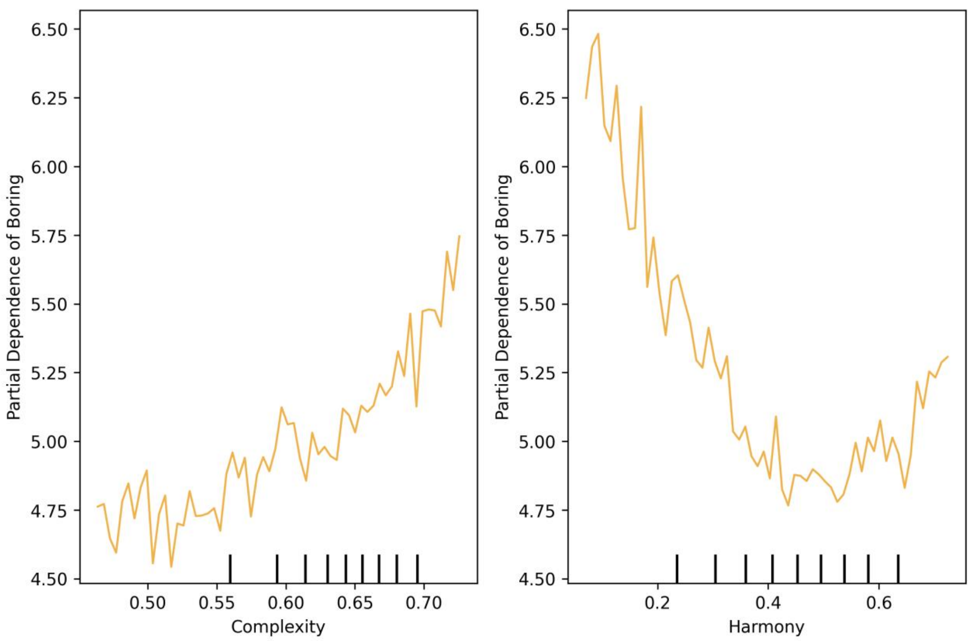
Taken together, these plots show that the effects of Cc and C_H are not well described by simple linear relationships. Instead, they often look like non-linear responses with thresholds and potential "sweet spots." Across outcomes, C_H shows stronger shape changes. Cc shows its clearest patterns as a monotonic increase for boring and a high-end decline for safe. These results provide useful cues for the next modeling stage. That stage will include multivariate controls and spatial structure.



(b) Safe



(c) Boring



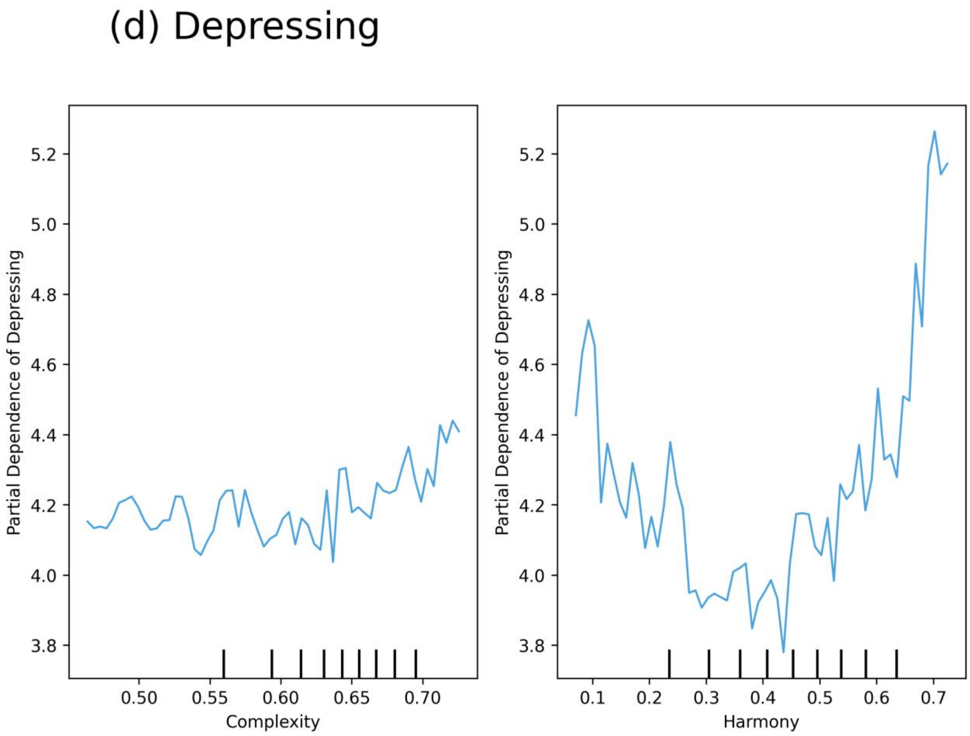
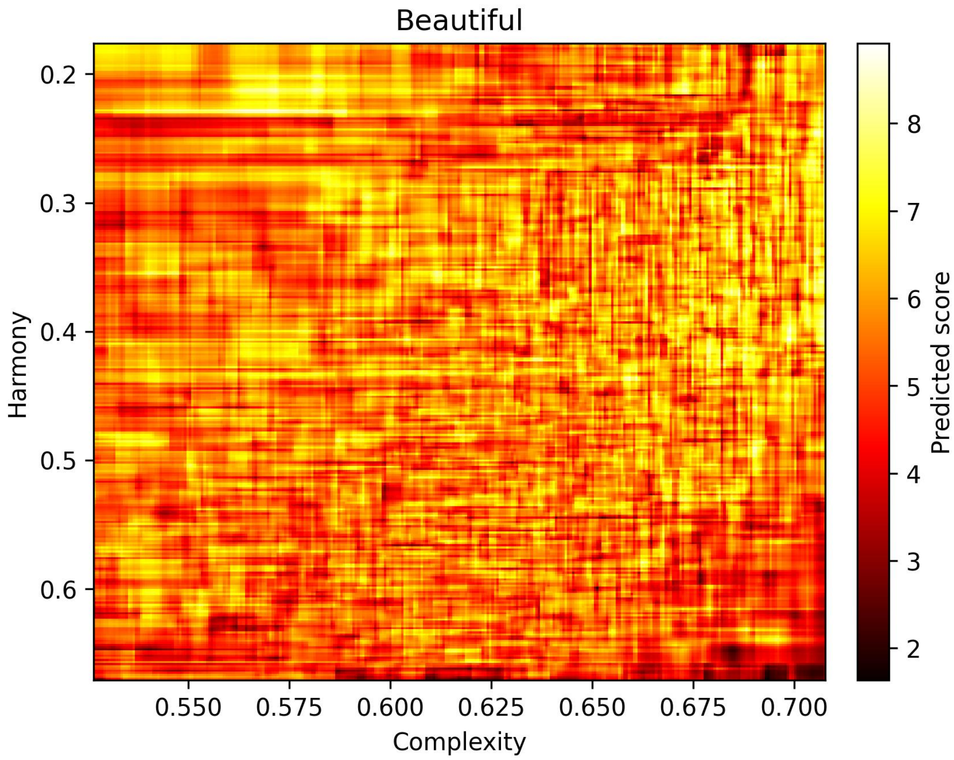


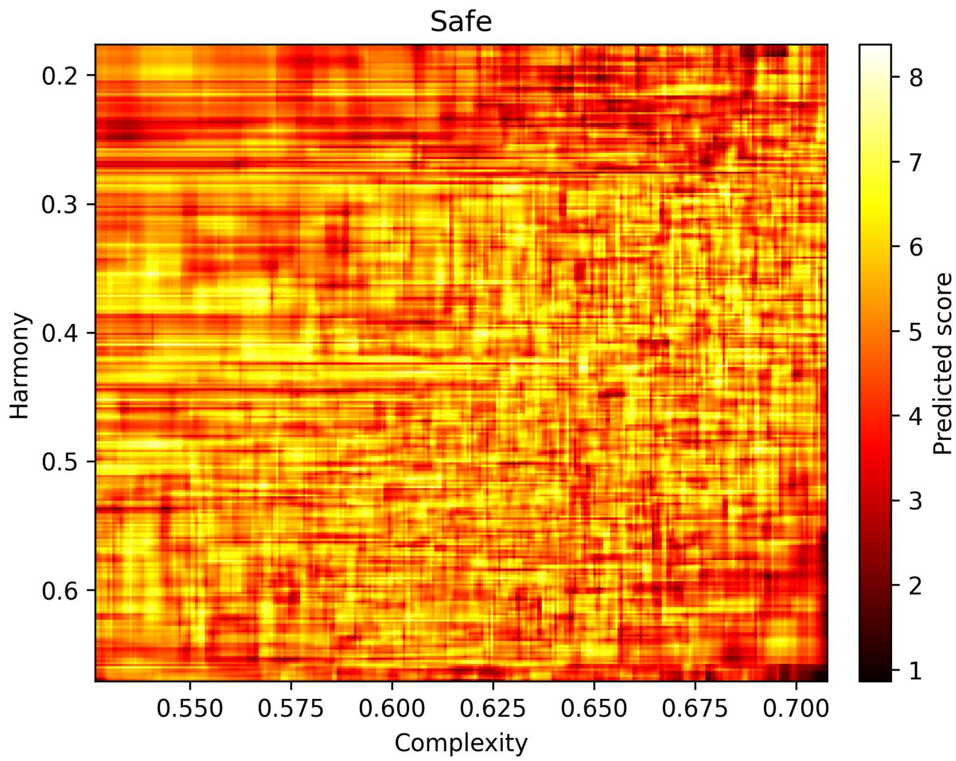
Figure 31. Bias dependence of color complexity and harmony on perception

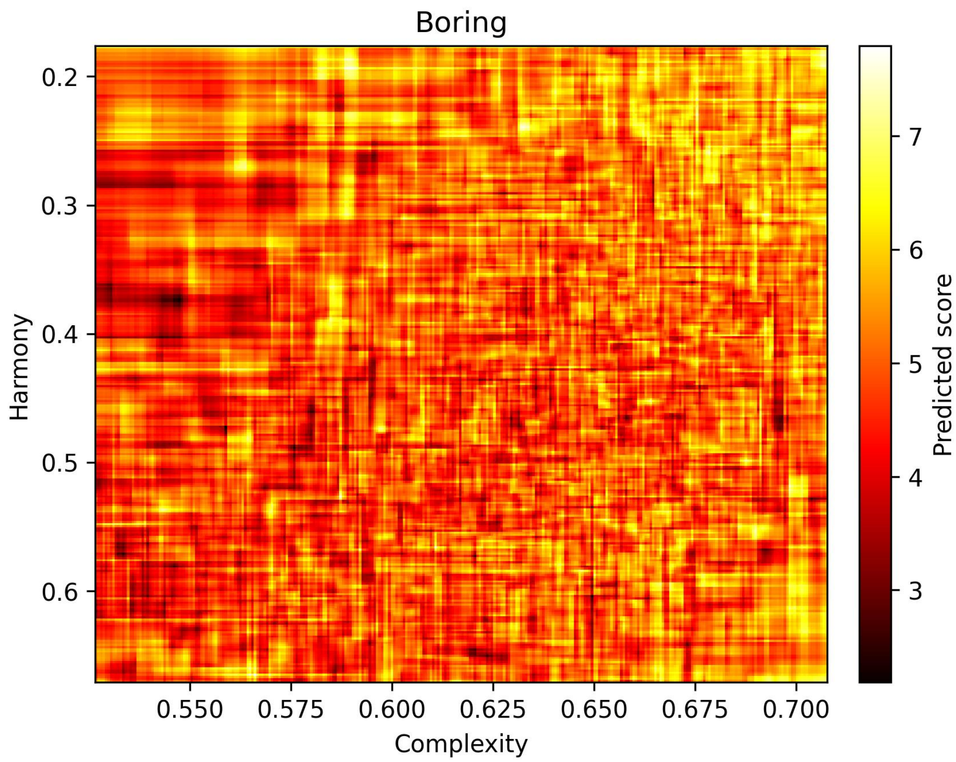
Then, figure 32 shows that color complexity (C_C) and color harmony (C_H) jointly shape emotional perception. These two metrics interact: some C_C-C_H combinations reinforce each other, while others weaken the predicted emotional response.

For Beautiful and Safe, the surfaces show a clear “top-left” concentration. The highest predicted values cluster where harmony is high. Complexity is low to moderate in this region. Given these positive perceptions, a high-harmony/moderate-complexity setup is the most favorable. The opposite corner—high complexity with low harmony—is consistently the darkest. This suggests that when the visual field is busy, but the palette lacks coherence, the model sees visual disorder. As a result, it strongly reduces predicted beauty and safety. In Zurich, this pattern is consistent with the idea that positive impressions are more likely in streetscapes that feel ordered and chromatically coherent than in scenes that look fragmented or cluttered.

For Boring and Depressing, the pattern flips. The highest predicted values concentrate in the “bottom-right” region, where complexity is high, and harmony is low. This indicates an amplifying configuration: visually crowded scenes with discordant color organization are not interpreted as “stimulating.” Instead, they are more likely to create cognitive and visual strain, which, in turn, raises predicted boredom and depressive scores. Consistent with this, the upper-left region—where Beautiful and Safe peak—stays relatively low on the surfaces for these two negative emotions.




Safe
Harmony
0.2
0.3
0.4
0.5
0.6
Complexity
0.550
0.575
0.600
0.625
0.650
0.675
0.700
Predicted score
1
2
3
4
5
6
7
8


Boring
Harmony
0.2
0.3
0.4
0.5
0.6
Complexity
0.550
0.575
0.600
0.625
0.650
0.675
0.700
Predicted score
3
4
5
6
7

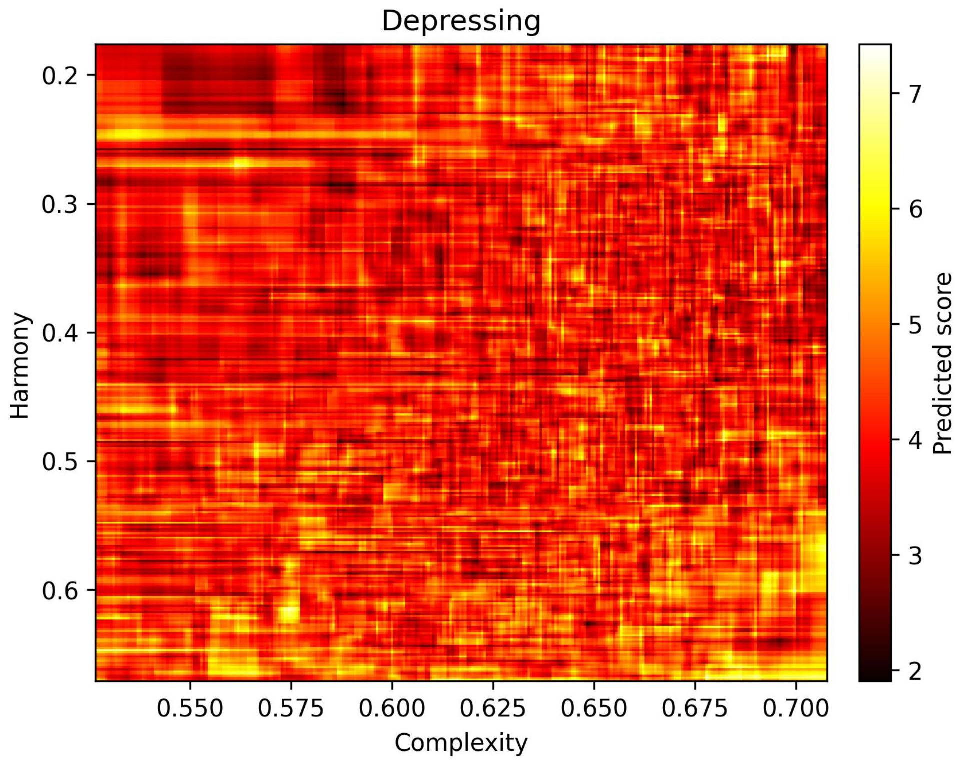


Figure 32. Color complexity-harmony interaction on perception

4.4.2 OLS and Geographically Weighted Regression

To compare a global linear model with a local model that can capture spatial heterogeneity in the color-emotion relationship, both Ordinary Least Squares (OLS) regression and Geographically Weighted Regression (GWR) were fitted. OLS treats the effects of color complexity and color harmony as constant across the study area. GWR, in contrast, allows coefficients to vary by location, which makes it better suited for testing spatial non-stationarity in street environments. In both approaches, color complexity and color harmony are used as predictors, and separate models are estimated for the four perception dimensions—Beautiful, Safe, Boring, and Depressing. Results are reported in Table 4.

In terms of model diagnostics and goodness-of-fit, OLS shows limited explanatory power: the adjusted R^2 values for the four emotions fluctuate only within the range of 0.00-0.04. Meanwhile, the Koenker (BP) test is significant in all four emotion models ($p < 0.001$), indicating pronounced non-constant residual variance, so a single set of global coefficients is unlikely to adequately represent how different neighborhoods respond to color features. In addition, the Jarque-Bera test is also significant for all emotion models ($p < 0.001$), suggesting that OLS residuals deviate from normality and further highlighting the limitations of a global linear framework in capturing the underlying data structure. After spatial non-stationarity is introduced, GWR delivers a clear improvement in model fit for all four emotions, with a substantial gain over OLS. At the same time, GWR yields markedly lower AICc values than OLS, implying that—even after penalizing model complexity—a locally varying coefficient model achieves a better overall fit. Regarding bandwidth, this study adopts an adaptive neighborhood with a bisquare kernel and selects bandwidth by minimizing AICc; the effective neighborhood size differs across emotions, which in turn suggests that the spatial processes underlying different perception dimensions operate at different smoothing scales. For Depressing, color complexity is not significant in the OLS model, while color harmony remains significant. This points to a clear difference between the two color measures. Depressive responses appear to be more closely tied to how well colors fit together than to visual complexity on its own. The spatial diagnostics add another layer to this result. The Koenker test is significant, and model fit improves noticeably under GWR. In other words, the effect of color complexity on depressing perception is not constant across space. It changes from one part of the city to another. For that reason, the non-significant OLS result should not be read as evidence that complexity does not matter. It is better understood as a sign that its effect is spatially contingent rather than globally uniform.

Table 4. OLS and GWR regression analysis

		OLS						GRW	
		Ratio	**P**	**Koenker**	**AdjR^2**	**AICc**	**Jarque-Bera**	**AdjR^2**	**AICc**
Beautiful	**C_C**	1.593017	0.000 *	0.000 *	0.01	52224.78	0.000 *	0.56	45906.84
	C_H	-1.403243	0.000 *						
Safe	**C_C**	-0.92084	0.021 *	0.000 *	0.00	53116.06	0.000 *	0.56	46344.6
	C_H	0.642295	0.000 *						
Boring	**C_C**	3.202793	0.000 *	0.000 *	0.04	45112.99	0.000 *	0.43	40336.69
	C_H	-1.383178	0.000 *						
Depressing	**C_C**	0.23106	0.406	0.000 *	0.01	44473.56	0.000 *	0.39	40639.31
	C_H	0.958116	0.000 *						

Note. *: $p < 0.05$

Further evidence for the complex relationship between street color and perception is provided by the distributional characteristics of the GWR local regression coefficients (Table 5). Across all emotion dimensions, the standard deviations of the local coefficients are substantially larger than the magnitudes of their corresponding means. Often, these deviations reach three to four times the mean. This indicates that the "color-emotion" relationship is not a uniform effect with a consistent direction. Rather, it is highly sensitive to Zurich's local urban contexts.

For Beautiful, the mean coefficient of complexity is positive (0.121). In contrast, harmony is negative (-0.187). The large standard deviations (0.473 and 0.627, respectively) suggest that, though complexity usually increases perceived beauty, its effect—and that of harmony—can be reversed by location. For Safe, the mean effect of complexity is positive (0.078), while the effect of harmony is close to zero (-0.009). Both effects show substantial dispersion (standard deviations above 0.48). Thus, perceived safety is not a universal response to color; it is likely moderated by specific street-interface characteristics, such as the contrast between active retail frontages and blank infrastructure walls.

For Boring, complexity and harmony have the smallest standard deviations in the dataset: 0.285 and 0.347, respectively. This suggests that boredom is likely the most stable perceptual response. Across neighborhoods, lower color complexity and monotonous harmony are associated with higher boredom scores. In contrast, Depressing shows a more distinctive pattern. The mean effect of complexity is negative (-0.051), while harmony is positive (0.162). Combined with large standard deviations, this indicates a pronounced local effect on depressive perception. Along high-traffic corridors, insufficient color diversity (negative C_C) may intensify stress. In other settings, infrastructure-induced "grey harmony" may become dominant. The strength—and even the direction—of these relationships varies with Zurich's landscape backgrounds and its road class hierarchy.

Table 5. Coefficient statistics of emotion-color GWR analysis

Emotional Perception	Color Complexity		Color Harmony	
	Average Value	**Standard Deviation**	**Average Value**	**Standard Deviation**
Beautiful	0.121	0.473	-0.187	0.627
Safe	0.078	0.487	-0.009	0.650
Boring	0.071	0.285	-0.070	0.347
Depressing	-0.051	0.362	0.162	0.480

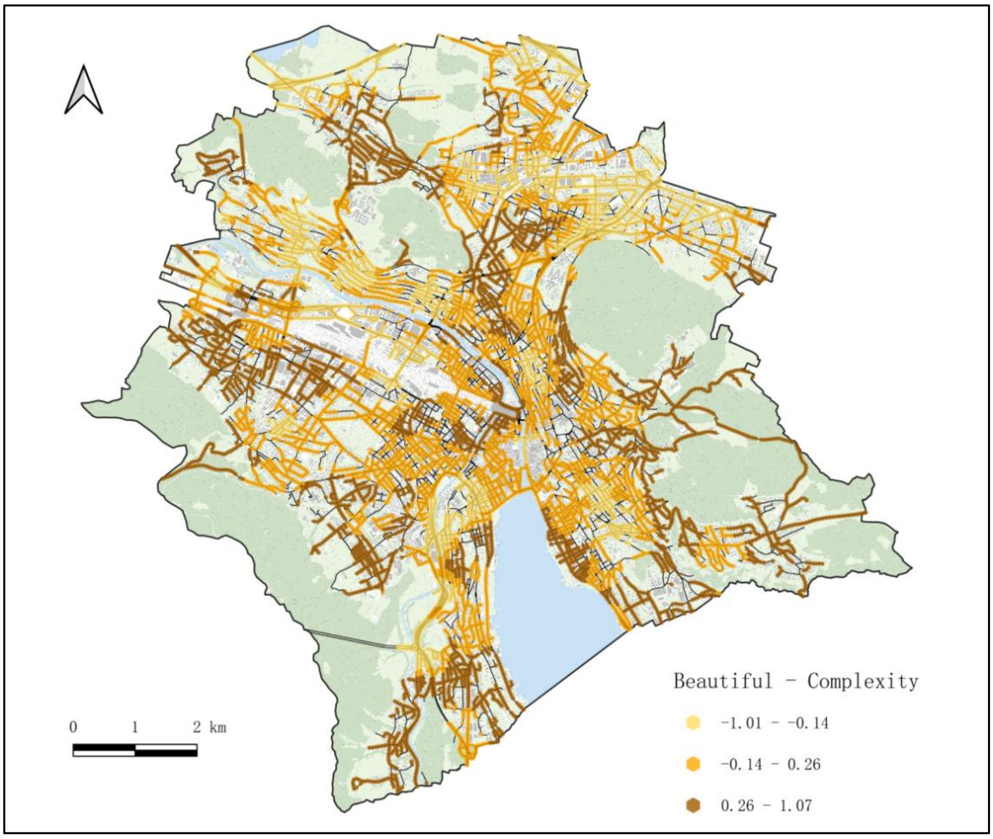
Overall, the comparison shows that OLS is useful as a basic model for average citywide effects, but the results indicate limits to relying on a single global linear approach. GWR fits the data better, and results show that the relationships between color complexity, color harmony, and emotional perception differ by location. These local changes provide stronger support for the next step: linking patterns to urban settings and possible reasons behind them.

4.4.3 Spatial Heterogeneity in the Color-Emotion Relationship

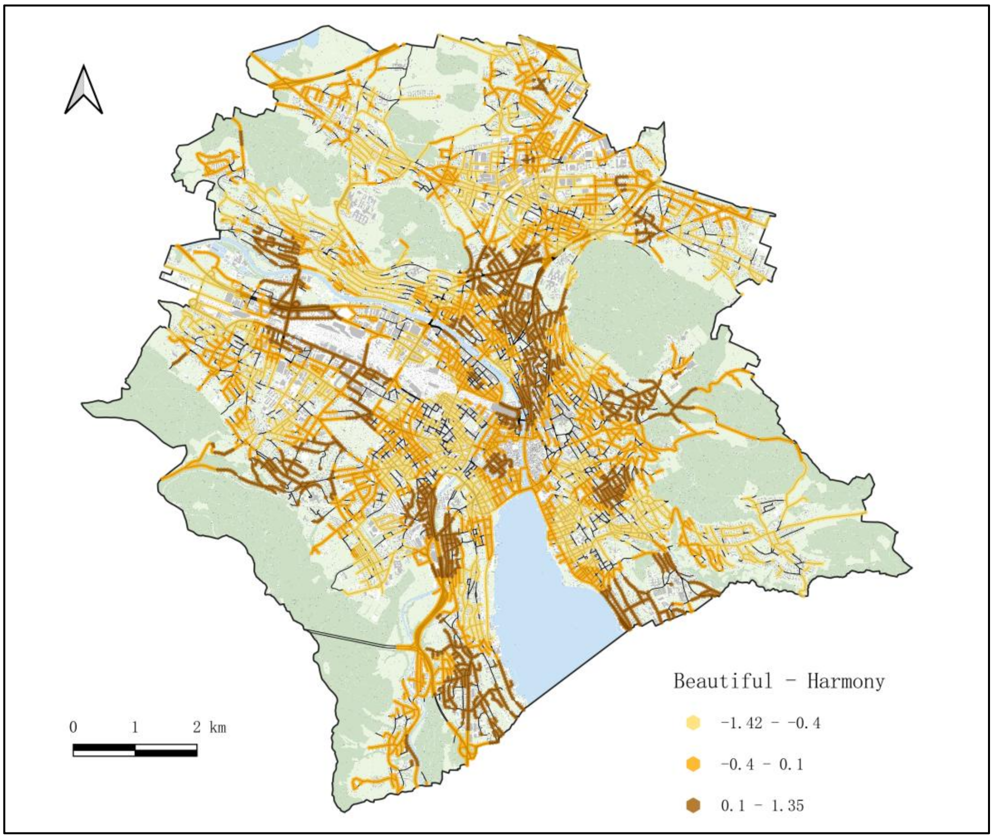
The following maps show how the color–emotion relationship changes across locations by mapping the local coefficients from the GWR models (Figures 33-36). Each pano_level segment is colored to reflect the size of its local coefficient. A positive coefficient means higher color values align with higher emotion scores, while a negative coefficient means higher color values align with lower scores. All coefficient maps use the same Natural Breaks classification, making it easier to see where local effects group or differ.

(1) The highest local coefficients for Beautiful-Complexity (GWR) appear in the west, where several corridor-like centers run through the city. In western Zurich, positive coefficients (dark brown) group together, creating continuous areas with high values. These positive chains also follow the main urban roads and corridors that connect the southern districts to the lake outlet. On busy streets with more varied and rich interfaces, higher color complexity often leads to higher beauty scores. From an environmental aesthetics point of view, complexity does not

always reduce appreciation. When it stays within a reasonable range, it can make scenes more interesting and encourage exploration. If the scene remains clear and ordered, richer colors and finer details can seem refined, lively, or modern, which increases perceived beauty. Weaker or negative coefficients (light yellow) are more common as broken segments in some outer areas and along roads to the east and northeast. In places that value calm, continuity, or a more consistent natural palette, extra complexity can feel like clutter or disturbance, lowering beauty ratings. This supports the idea that there is an optimal range for complexity and preference: if complexity goes beyond what people see as orderly and readable, its effect can change from positive to negative. For Beautiful-Harmony, the local coefficients show a stronger central axis, with corridor-like structures and several centers instead of a simple concentric pattern. Strong positive coefficients (dark brown) are found along the main road network and major connectors, stretching outward along several main corridors. In street environments where visual order, interface coherence, and spatial clarity are already strong, higher color harmony is more likely to increase beauty scores. Negative coefficients (light yellow) do not form clusters but show up as scattered patches in some outer areas and around certain nodes or corridors. This suggests that in places with sharper material contrasts, mixed functions, or more stimulating streetscapes, very high color consistency may stop adding to beauty and can even have a negative effect. It is important to remember that color-aesthetics research sees harmony and aesthetic preference as related but different, and their connection can change depending on the main elements in view, such as natural backdrops, building facades, and street infrastructure. In practice, weaker or mixed effects are more common in hillside residential areas and along some lakeside roads, rather than a consistent negative relationship.



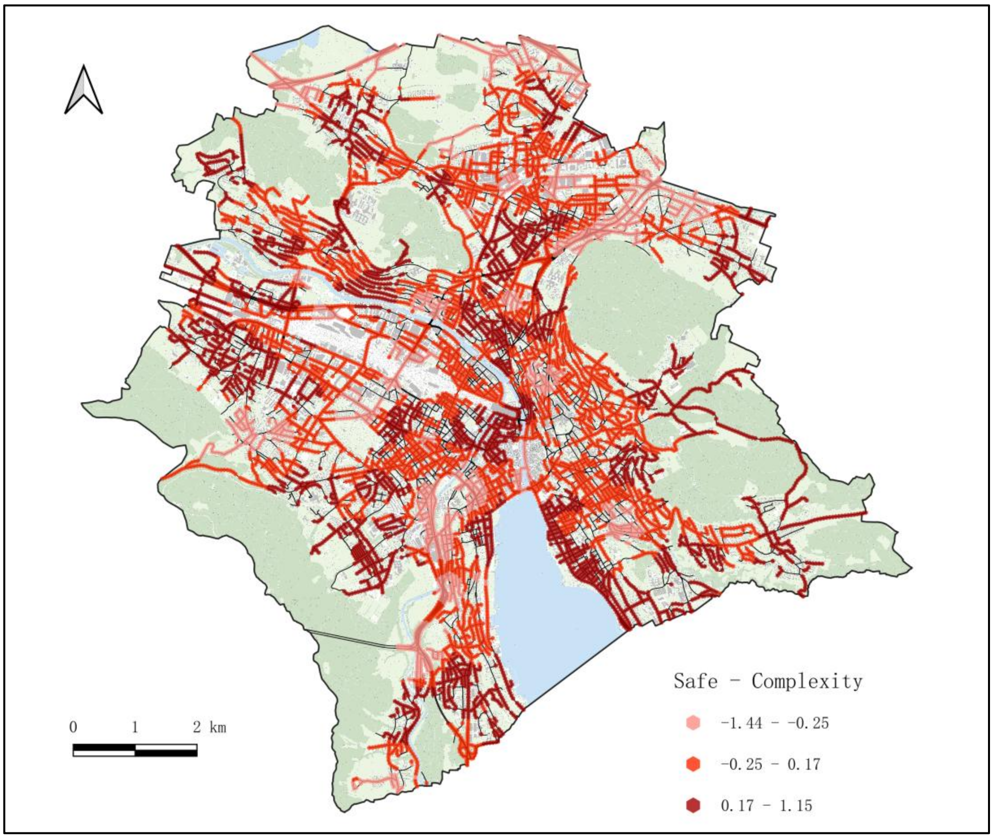
(a) Beautiful-Complexity



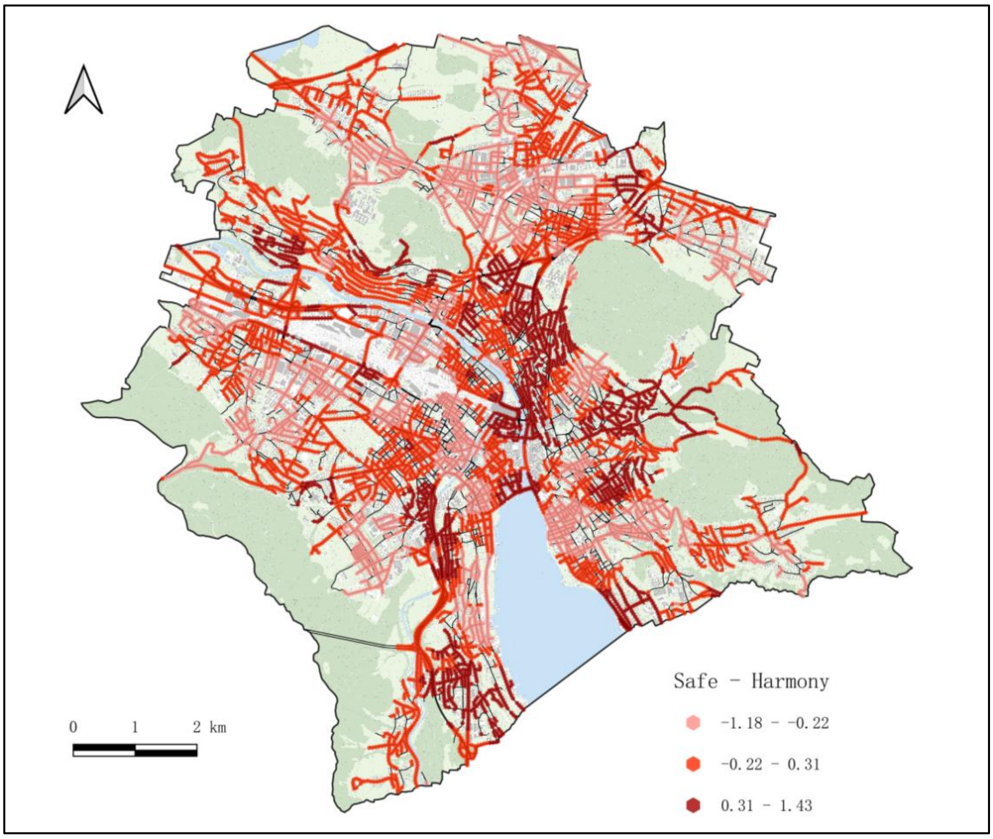
(b) Beautiful-Harmony

Figure 33. Visualization of GWR analysis of beautiful perception-colors

(2) For Safe-Complexity, many local effects are positive, especially in the city center and along main roads. In these places, more to look at often makes people feel safer. This fits the idea of “active street-natural surveillance.” Here, the complexity comes from people being out in different kinds of places mixed together. Negative effects are less common and mostly occur in outer areas, places with lots of traffic, or at edges, such as near big parks or busy crossroads. Here, the complexity comes from guardrails, road crossings, signs, and other road features. These can overwhelm people and make them feel less safe. For Safe-Harmony, the biggest positive effects are on the main urban roads and often stretch for long distances. On business streets or new mixed-use roads with neat street edges, matched colors are more likely to make people feel safe. But negative effects turn up in some northern and mid-outer areas. This means the link between harmony and safety changes in different places. On roads crowded with transport or at busy crossings—with less order—harmony does not always make people feel safer. Sometimes its effect can fade or become negative.



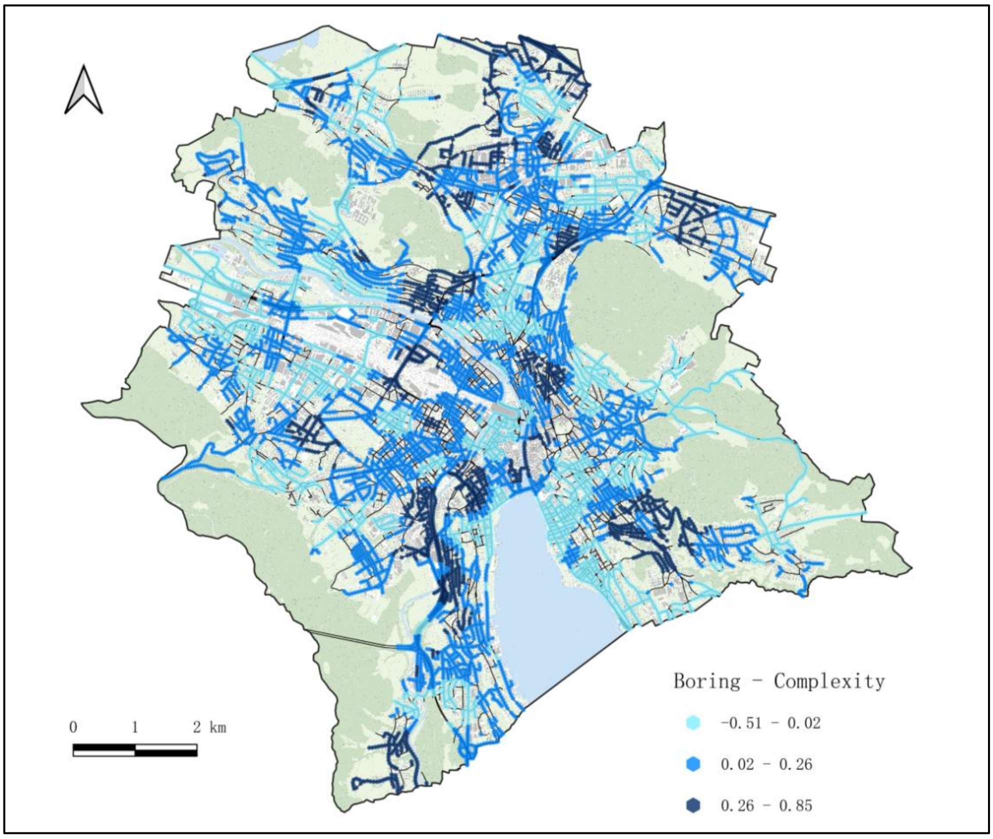
(a) Safe-Complexity



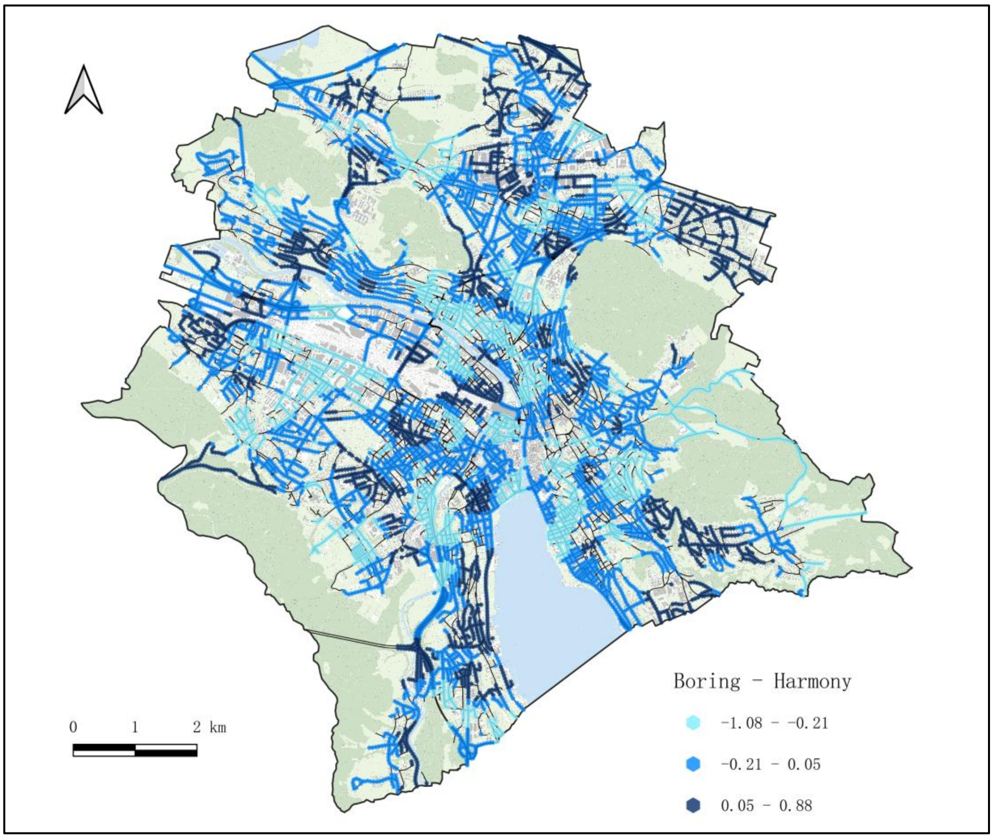
(b) Safe-Harmony

Figure 34. Visualization of GWR analysis of safe perception-color

(3) The local coefficient results for Boring-Complexity show that more complexity usually leads to greater boredom. This is especially true along main roads and in areas with heavy traffic. Complexity here often comes from layers of streets, signs, and features that break up the view. This type of complexity is different from the details that hold your attention. On some inner-city streets, things are more stable and easier to understand. In these places, more complexity can actually make things more interesting and reduce boredom. This suggests boredom does not depend only on how much is happening. It also depends on whether the setting gives useful information at the right level. This helps explain why complexity can affect each street differently. For Boring-Harmony, the results are mixed. In some areas, especially along main roads and near intersections, higher harmony does not always lower boredom. Instead, it can make things look too similar and predictable. This increases boredom. On some inner streets, where it feels more like a neighborhood, a more orderly look and matching colors can help. They make the area easier to view and less tiring, which reduces boredom. Overall, these findings suggest that harmony does not always affect boredom the same way. Its effect depends on a balance between order and new things. When scenes are too similar or lack interesting elements, people are more likely to get bored.



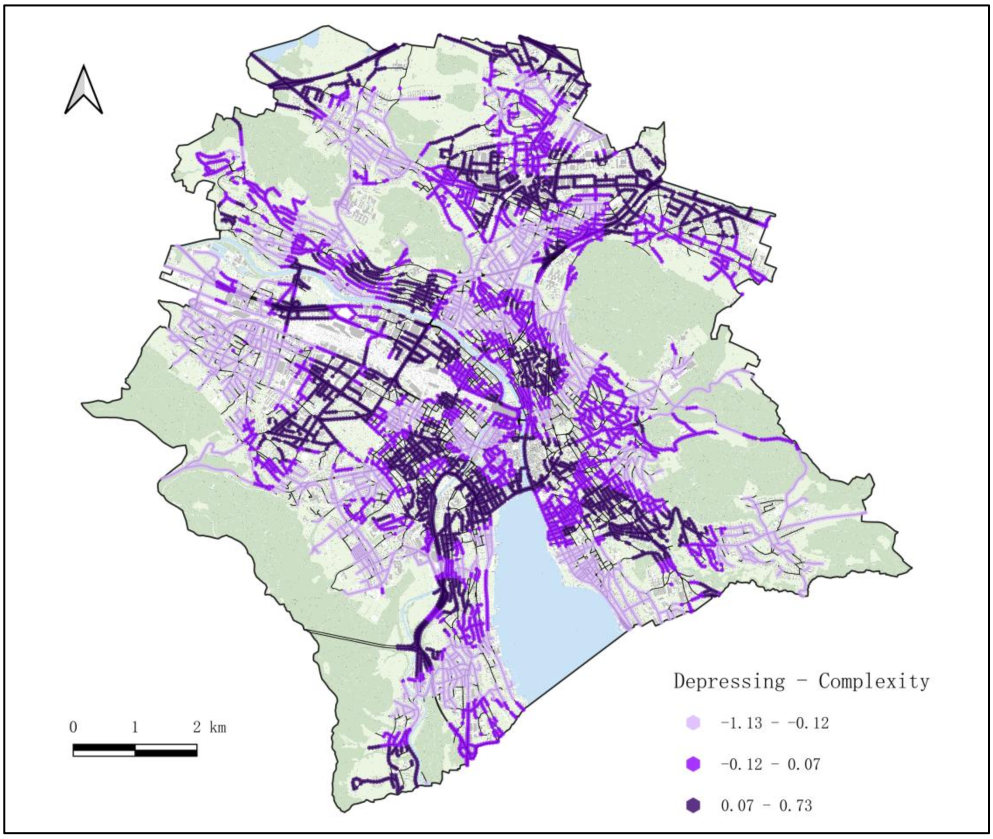
(a) Boring-Complexity



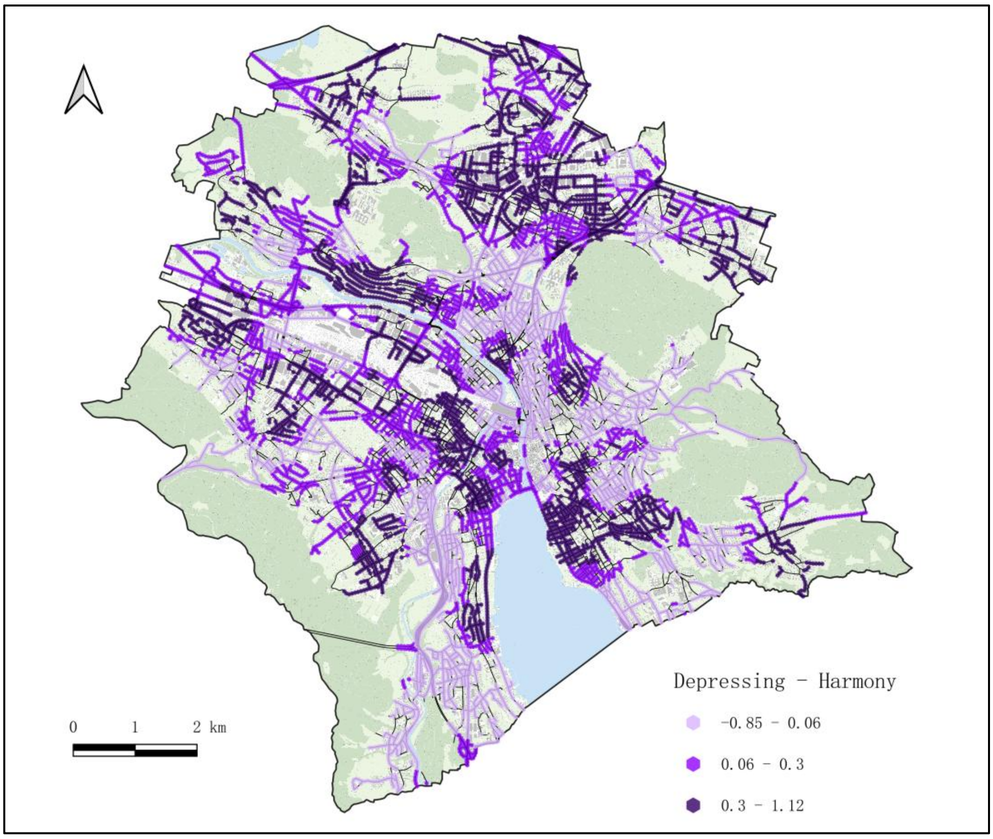
(b) Boring-Harmony

Figure 35. Visualization of GWR analysis of boring perception-colors

(4) In the Depressing-Complexity maps, high values (dark purple) are most evident in city centers and spread along main roads. This creates a pattern with a busy center and lines of complexity stretching outward. On busy streets with heavy traffic and lots of features, extra complexity can feel like clutter and too much information, for example, frequent signs, packed facilities, and many traffic features. This means higher complexity can lead to more depression. This fits the urban overload theory, which holds that excessive complexity can be a problem rather than an interesting challenge. Lower values (light purple) appear on some roads outside the center and nearby side streets. In areas with more stable settings, more natural views, or fewer distractions, extra complexity does not always cause stress. Its effect on depression can be weaker or even helpful, which shows that the effect depends on location. For Depressing-Harmony, the pattern depends more on context. Strong high values (dark purple) appear in long or partial stretches along some main roads and at certain intersections, rather than being common by the lake or in the south residential areas. At the same time, low values and weak effects (light purple and bright purple) are common, suggesting that greater color harmony does not always increase depression, and that its effect varies by location. Other studies often link higher harmony with calmer, more relaxing places. However, on busy streets under greater stress, harmony can come at the cost of more sameness, as well as crowding and noise. This may help explain why positive links with negative feelings still show up in some places.



(a) Depressing-Complexity



(b) Depressing-Harmony

Figure 36. Visualization of GWR analysis of depressing perception-colors

4.5 Hierarchical Multiple Regression: Net Effects of Color Index

To test whether color complexity (C_C) and color harmony (C_H) still show independent associations with the emotion scores after accounting for streetscape semantic structure, a stepwise multiple regression setup was used, with four nested models (Figure 37):

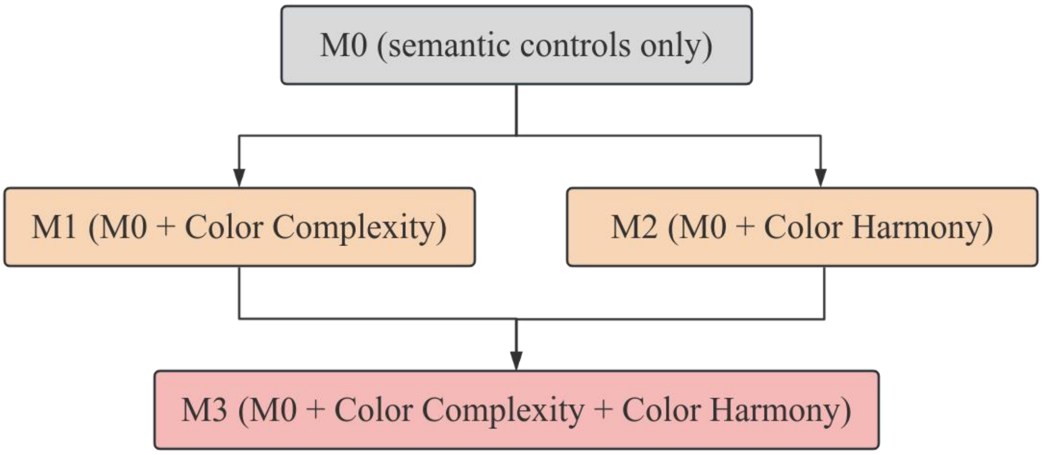


Figure 37. Hierarchical Multiple Regression Model Workflow

Semantic controls are the six scene components: road, sidewalk, building, wall, vegetation, and sky. To keep models clear and reduce overlap, for each emotion, the M0 control set is picked separately using AIC-based forward selection. After fitting these initial models, adjusted R^2 and AIC/BIC are used to assess model fit and to demonstrate both improvements and the impact of added complexity. Transitioning to joint models, specifically in M3, a nested F-test checks whether adding color variables significantly improves fit. Throughout, all coefficients use HC3 robust standard errors to address potential heteroskedasticity.

4.5.1 Semantic Baseline (M0)

As shown in Table 6, the M0 baseline model with semantic controls alone already provides substantial explanatory power. This indicates that streetscape semantic composition accounts for a meaningful share of variation in the emotion scores. As a result, the main question in the following models is not whether semantics matter. The key test is whether, after controlling for semantics in M0, adding the color variables further increases adjusted R^2, lowers AIC/BIC, and yields color coefficients that remain statistically significant.

Table 6. M0 Model Fit Results

Emotional Perception	Adj. R^2	AIC	BIC	M0 semantic controls
Beautiful	0.516	43775.74	43820.03	vegetation, building, road, sidewalk, sky
Safe	0.404	47013.60	47057.88	vegetation, building, road, sidewalk, sky
Boring	0.222	42590.52	42634.81	vegetation, building, road, sidewalk, sky
Depressing	0.279	40694.88	40746.54	vegetation, building, road, sidewalk, sky,wall

4.5.2 Adding Color Complexity (M1 = M0 + Color Complexity)

Once C_C is added to the semantic baseline (M1), the degree of improvement across outcomes differs. This is most evident when comparing adjusted R^2 to AIC/BIC.

For Safe and Boring, Table 7 shows a clear improvement from M0 to M1: adjusted R^2 increases, and both AIC and BIC drop. The C_C coefficient is also robustly significant ($C_{C}\_p < 0.05$). Taken together, this suggests that after semantic composition is controlled, C_C still contributes independent information for safety and boredom—but in opposite directions (higher complexity predicts lower safety and higher boredom).

For Beautiful, C_C remains statistically significant, but the gain is small and sensitive to model penalties: ΔAdj. R^2 is only 0.000188, AIC falls slightly, while BIC increases. This points to a net effect that exists but is marginal, and not strongly supported once complexity is more heavily penalized.

For Depressing, adding C_C alone does not produce meaningful improvement: AIC is essentially unchanged, BIC worsens, and the C_C coefficient is not significant. In other words, depressing scores do not appear to be reliably explained by complexity alone in M1, which suggests that any color-related mechanism for depressing is unlikely to be captured by a single complexity term.

Table 7. M1 Model Fit Results

Emotional Perception	Adj. R^2 (M1)	AIC (M1)	BIC (M1)	ΔAdj. R^2 (M1–M0)	ΔAIC (M1–M0)	ΔBIC (M1–M0)	C_C_β	C_C_p
Beautiful	0.517	43772.13	43823.79	0.000188	-3.62	3.76	-0.623	2.26e-02
Safe	0.405	46994.56	47046.23	0.001006	-19.03	-11.65	-1.381	8.82e-06
Boring	0.226	42524.72	42576.38	0.004372	-65.81	-58.43	2.055	2.20e-15
Depressing	0.279	40694.79	40753.83	0.000067	-0.09	7.29	0.334	1.60e-01

4.5.3 Adding Color Harmony (M2 = M0 + Color Harmony)

Compared with complexity, C_H shows a more outcome-specific pattern in M2, as reflected in the adjusted R^2 and the information criteria (Table 8).

For Beautiful and Depressing, moving from M0 to M2 produces a clear increase in adjusted R^2, while AIC and BIC both decrease, and the C_H coefficient is highly significant ($C_H\_p < 0.05$). This consistent shift suggests that even when streets have similar semantic composition, differences in how colors are arranged and combined still matter for overall judgments of beauty and depressingness. For Boring, the improvement is modest. Some fit indicators move in the right direction, but the change is much smaller than for the two outcomes above, which suggests harmony adds only limited information for boredom. For Safe, M2 provides little to no improvement: adjusted R^2 barely changes, AIC/BIC do not improve, and the C_H coefficient is not significant. This implies that once semantics are controlled, perceived safety is more strongly tied to structural and visibility-related cues than to color coordination alone.

Table 8. M2 Model Fit Results

Emotional Perception	Adj. R^2 (M2)	AIC (M2)	BIC (M2)	ΔAdj. R^2 (M2–M0)	ΔAIC (M2–M0)	ΔBIC (M2–M0)	C_H_β	C_H_p
Beautiful	0.520	43689.64	43741.31	0.003540	-86.10	-78.72	-0.916	2.05e−19
Safe	0.404	47015.49	47067.15	-0.000045	1.89	9.27	-0.037	7.37e−01
Boring	0.223	42578.22	42629.88	0.000873	-12.31	-4.93	-0.352	1.35e−04
Depressing	0.290	40505.09	40564.14	0.011513	-189.79	-182.41	1.186	2.24e−40

4.5.4 Joint Model (M3 = M0 + Color Complexity + Color Harmony)

When C_C and color harmony C_H are added together (M3), the most consistent improvement appears across all four emotion outcomes (Table 9). Relative to M0, M3 increases adjusted R^2 for every outcome (ΔAdj. $R^2 > 0$), and both AIC and BIC drop throughout (ΔAIC < 0; ΔBIC < 0). The nested F-test is also significant in all four cases, rejecting the null that the color block adds no explanatory value. At the model level, this indicates that once semantic structure is controlled, the color variables as a set still provide meaningful additional explanatory power for perceived emotions.

The coefficient patterns in M3 also suggest that the two color indicators play different roles. C_C remains significant for all four emotions ($C_C\_p < 0.05$), pointing to a broad and stable contribution. C_H, in contrast, is strongly significant for Beautiful and Depressing ($C_H\_p < 0.05$), not significant for Boring, and only marginal for Safe ($C_H\_p \approx 0.072$). This pattern implies that harmony acts more like a selective shaper of overall atmosphere or emotional tone, rather than a uniform predictor across all perception dimensions.

Table 9. M3 Model Fit Results

Emotional Perception	Adj. R^2 (M3)	AIC (M3)	BIC (M3)	ΔAdj. R^2 (M3–M0)	ΔAIC (M3–M0)	ΔBIC (M3–M0)	F-test_p	C_C_β	C_C_p	C_H_β	C_H_p
Beautiful	0.521	43661.90	43720.94	0.004702	-113.85	-99.08	2.68e-26	-1.496	1.98e−07	-1.083	2.15e-24
Safe	0.405	46993.36	47052.40	0.001117	-20.25	-5.48	5.50e-06	-1.550	2.31e−06	-0.210	7.17e-02
Boring	0.226	42524.80	42583.85	0.004432	-65.72	-50.96	7.43e-16	1.947	9.47e−13	-0.134	1.65e-01
Depressing	0.292	40472.48	40538.91	0.013522	-222.40	-207.64	7.52e-50	1.411	1.67e−08	1.342	3.49e-47

The stepwise sequence from M0 to M3 leads to a consistent takeaway. Streetscape semantics explain much of the variation in emotion scores. Yet, the coordinated improvements in adjusted R^2, AIC/BIC, and nested tests show that color still matters after accounting for semantics. In other words, color complexity and color harmony form an independent, identifiable explanatory layer in street-level emotional perception.

4.6 Cross-City Validation: Testing Robustness and Transferability

To test whether the proposed street color-affective perception framework is robust and transferable across urban contexts, a cross-city replication was carried out in Geneva in addition to Zurich. The purpose of this replication is not to reproduce every indicator and module at full scale. Instead, given computational and practical constraints, a reproducible-sample strategy is used to check a small set of key methodological "checkpoints" that capture the main causal logic of the framework. Two mechanism-level conclusions are treated as core targets: (i) color metrics still add incremental explanatory power after controlling for streetscape semantics (the color net effect); and (ii) color complexity (C_C) and color harmony (C_H) show different effect profiles across emotions. If these two signals remain directionally consistent and statistically supported in a second city, then even a lightweight replication provides meaningful evidence for the framework's external validity.

4.6.1 Purpose of Cross-City Replication and City Selection

To test if the core findings hold elsewhere, a small-scale replication was done in a city outside Zurich. Rather than rebuilding the whole pipeline at the same scale, this replication uses reproducible samples and an isomorphic modeling setup. The main question is whether the two mechanism-level conclusions point in the same direction and remain statistically significant in a different urban context.

Geneva was chosen for three main reasons:

(1) Contextual contrast with comparability: Zurich and Geneva are both major Swiss cities, with similar data access and infrastructure conditions. At the same time, Geneva is in the French-speaking region and differs in urban character and spatial structure. That makes it a stronger test of cross-context transferability rather than a "same-distribution" rerun, which strengthens the external validity argument.

(2) Varied streetscape settings. Geneva features several common urban settings: lakefront streets, historic districts, and dense central areas. This allows key street types to be covered even with a smaller sample.

(3) Low transfer cost for the core variables. The main variables (perception scores, semantic proportion controls, and the two color metrics C_C and C_H) depend little on city-specific auxiliary datasets. Thus, the same indicator definitions and model specification can be reused as in Zurich. This supports a clearer cross-city interpretation.

For the Geneva replication, street view images were obtained from the Google Street View API, the road network was derived from OpenStreetMap, and the administrative boundary was based on the CAD_COMMUNE dataset of Geneva commune boundaries. After running the same processing workflow as in the Zurich analysis, 5,728 images were retained for the Geneva analysis. These correspond to 2,864 unique pano_ids.

4.6.2 Replication Results: Color Net Effect

Using the same M0→M3 hierarchical setup as Zurich, Geneva tests whether adding the color block (C_C + C_H) continues to improve model fit after controlling for semantic composition. Table 10 shows that this color effect is reproduced in Geneva, though it varies by outcome.

Negative emotion dimensions show the clearest support. Depressing has the greatest improvement, largest increase in adjusted R^2, largest decreases in AIC and BIC, and a highly significant F-test comparing M3 with M0 (*F-test_p* < 0.05). Here, C_H is the main driver, while C_C is not significant. Boring also improves across all metrics; the significant predictor is C_C, not C_H. For Beautiful, the evidence is weaker but present: the nested comparison is significant (*F-test_p* < 0.05), AIC decreases slightly, adjusted R^2 increases slightly, and BIC rises. At the coefficient level, C_C is significant, whereas C_H is only marginally significant. Safe shows no meaningful improvement, with little change in adjusted R^2, worse AIC and BIC, a non-significant F-test, and no significant color coefficient.

Overall, the Geneva replication confirms that street color can retain independent explanatory value beyond semantic composition, especially for Boring and Depressing, but not uniformly across all emotional perception dimensions.

Table 10. M3 Model Fit Results (Geneva)

Emotional Perception	Adj. R^2 (M3)	AIC (M3)	BIC (M3)	ΔAdj. R^2 (M3–M0)	ΔAIC (M3–M0)	ΔBIC (M3–M0)	F-test_p	C_C_β	C_C_p	C_H_β	C_H_p
Beautiful	0.443	10568.51	10604.27	0.001075	-3.53	8.39	2.34e-2	-1.207	3.39e−02	-0.422	5.27e−02
Safe	0.238	11362.40	11410.08	0.000242	1.09	13.00	2.34e-1	-0.583	3.56e-1	-0.380	1.29e-1
Boring	0.113	10186.39	10228.11	0.007496	-22.10	-10.18	2.22e-06	2.495	1.12e-06	-0.045	8.20e-1
Depressing	0.183	9761.48	9809.16	0.016714	-55.97	-44.05	1.038e-13	0.233	6.38e-1	1.397	1.29e-13

4.6.3 Correlation Analysis between Streetscape Color and Emotional Perception (Geneva)

Table 11 reports Pearson correlations between the two color indicators and the four emotion scores in the Geneva sample, where r denotes the Pearson correlation coefficient. Overall, the sign pattern largely matches what is observed in Zurich. This suggests the color-emotion link is unlikely to be a one-off, city-specific artifact and is more consistent with a transferable tendency.

In Geneva, C_C is positively correlated with Beautiful ($r = 0.054$, $p < 0.01$) and Boring ($r = 0.105$, $p < 0.01$), and negatively correlated with Depressing ($r = -0.056$, $p < 0.01$), while its correlation with Safe is close to zero and not statistically significant ($r =$

0.011, $p > 0.01$). The directions are the same as in Zurich for the significant associations. In this bivariate view, higher color richness/diversity tends to co-occur with stronger visual stimulation: it is linked not only to slightly higher beauty, but also more clearly to boredom-related variation, while showing only a weak relationship with safety in Geneva. C_H shows a negative correlation with Beautiful ($r = -0.208$, $p < 0.01$), Safe ($r = -0.058$, $p < 0.01$), and Boring ($r = -0.136$, $p < 0.01$), and a positive correlation with Depressing ($r = 0.207$, $p < 0.01$). These directions also align with Zurich, which strengthens the cross-city consistency claim.

Correlations do, however, vary in strength across outcomes in Geneva. For C_C, the association with Safe is essentially absent, and the correlation with Beautiful is also relatively small compared with those for Boring and Depressing. By contrast, C_H-Beautiful and C_H-Depressing stand out as the strongest bivariate associations in Geneva, whereas the link with Safe is present but much weaker. This pattern is not unexpected because these are bivariate correlations without semantic controls. Perceptual judgments—especially those related to safety and beauty—are often shaped jointly by color and non-color cues, such as enclosure, activity, built-form structure, and traffic conditions. When scene composition is not held constant, some color–emotion associations may appear weaker, whereas others remain comparatively pronounced.

Taken together, Table 11 supports cross-city validity: the color-emotion relationship transfers in the same direction across most dimensions, while still showing clear outcome-specific differences.

Table 11. Color-emotion perception correlation analysis (Geneva)

Color Index	Beautiful	Safe	Boring	Depressing
Color Complexity	0.054 **	0.011	0.105 **	-0.056 **
Color Harmony	-0.208 **	-0.058 **	-0.136 **	0.207 **

Note. **: $p < 0.01$

4.6.4 Coefficient Statistics of the Emotion-Color GWR Analysis (Geneva)

To test whether the spatially varying color–emotion relationships observed in Zurich also appear in a different urban setting, the Geneva GWR results are summarized using the mean and standard deviation of the local coefficients (Table 12). In a local modeling context, a key replication signal is a stable sign structure across cities. In Geneva, the mean Cc coefficient is positive for Beautiful (0.066), Safe (0.077), and Boring (0.101), and negative for Depressing (-0.034). The mean C_H coefficient shows the opposite pattern: it is negative for Beautiful (-0.212), Safe (-0.016), and Boring (-0.097), but positive for Depressing (0.227). This sign configuration matches Zurich, which suggests the framework is picking up a robust, transferable structure rather than a city-specific artifact—even when coefficients are allowed to vary by location.

At the same time, coefficient dispersion in Geneva is smaller than in Zurich, which indicates weaker spatial spread in the local relationships. A plausible reason is Geneva's more compact study area and more concentrated urban form. Under the same GWR specification, lower dispersion is therefore expected. This pattern is better read as a more coherent spatial expression of the color–emotion linkage, not as evidence that the relationship disappears. Overall, Table 12 supports cross-city transferability in the core directional trends, while still leaving room for interpretable differences in effect size and spatial heterogeneity.

Table 12. Coefficient statistics of emotion-color GWR analysis (Geneva)

Emotional Perception	Color Complexity		Color Harmony	
	Average Value	Standard Deviation	Average Value	Standard Deviation
Beautiful	0.066	0.377	-0.212	0.415
Safe	0.077	0.384	-0.016	0.437
Boring	0.101	0.250	-0.097	0.258
Depressing	-0.034	0.228	0.227	0.341

5 Discussion

Reducing negative feelings and supporting positive everyday experiences make cities healthier and more livable. As urban data systems and analytics have matured, it is now practical to study these experiences at scale by linking emotional responses to the spatial patterns of many environmental features across a city. As cities keep growing, this study uses Zurich as a case area and combines large-scale street view imagery with deep learning–based perception prediction to measure multiple dimensions of street-level emotional perception. Focusing on urban color, it examines how street color complexity and color harmony relate to emotional perception, how these effects vary across space, and whether they remain visible after semantic scene composition is taken into account. Using spatial analyses in Zurich and a cross-city replication in Geneva, the results show that: (i) urban perception has a clear spatial structure and forms a mappable "perception geography"; (ii) color complexity and color harmony follow recognizable but different urban-geographic patterns; (iii) their relationships with emotional perception are not purely linear, but often outcome-specific, interactive, and spatially heterogeneous; and (iv) even in hierarchical multiple regression models with semantic controls, street color still retains a measurable net effect, although its strength and direction vary by outcome and local context. The Geneva replication further suggests that these core patterns—including the net effect after semantic control and the directional structure of local coefficients—are not limited to Zurich alone, but are consistent with a more transferable framework for studying color-emotion relationships in urban streets. The main conclusions are summarized below.

5.1 Discussion of Research Question 1: What spatial patterns characterize human emotional perceptions of street environments?

One key finding is that emotional perceptions in Zurich are not randomly distributed.

Tests for Safe, Beautiful, Boring, and Depressing show that similar perception levels tend to cluster along the street network. Positive perceptions, such as Safe and Beautiful, show the strongest clustering across the city. Negative perceptions, such as Boring and depressing, also form clusters, but these are more fragmented and often follow corridors. This suggests that negative feelings are often tied to specific transport and land-use areas, rather than following a simple center-to-edge pattern. These spatial patterns also reveal more about how different perceptions are related.

These four perception types also show a clear internal structure. Beautiful and Safe are closely linked, while Depressing usually moves in the opposite direction. Boring acts as a partly independent "monotony" factor: it is related to Depressing but is not just the opposite of Beautiful. This suggests that negative perceptions have several components, not just one. The hotspot analysis makes these patterns easier to see on the map. Positive hotspots are found on streets with better-quality surroundings, such as parts of the inner city and areas near water or green spaces. Negative hotspots are more common along busy infrastructure corridors and in outer transition zones. Overall, these results show that Zurich has a clear, mappable "perception geography."

5.2 Discussion of Research Question 2: What spatial distribution patterns do streetscape color complexity and color harmony exhibit?

The Zurich study shows that street colors vary by location. The two metrics highlight different "color geographies." Color complexity (C_C) is not only highest in the historic center. Higher C_C values are found along major roads, in the dense city network, and in some outer or transitional areas. These areas often have mixed uses and varied edges. This breaks up the view into more distinct color patches and creates stronger local contrasts. In contrast, lower to moderate C_C values appear as embedded, discontinuous patches across the network rather than forming a single stable, low-complexity zone. Parts of the old town and some central streets near the lake and river may appear visually coherent. However, these areas do not show uniform low complexity. Instead, they often contain a mix of medium- to high-segment users.

Color harmony (C_H) is usually higher in the city center and in visually unified areas, especially when buildings and materials are coordinated. Lower C_H is more common in areas with mixed uses or heavy infrastructure, where different materials, bold signs, and overlapping land uses make the main colors stand out more from each other.

These patterns do not suggest a simple "good vs. bad color" split. Instead, they show different qualities. Complexity means variety and broken-up edges. Harmony refers to how well the colors fit together. This brings us to the next question: how people view a "more complex" color setting likely depends on whether that complexity is accompanied by high harmony (rich and unified) or low harmony (jarring and discordant). Thus, the relationship between complexity and harmony shapes public perception of streetscapes.

5.3 Discussion of Research Question 3: How do color complexity and color harmony influence human emotional perceptions of streetscapes?

Street color does influence how people feel, but this impact is not strong when looking at just two variables. The effect also depends on which outcome is measured. Pearson correlations suggest many factors besides color matter. Even so, the findings show color should not be ignored. When the scene's content stays the same, the hierarchical M0-M3 tests show that color still helps explain results. However, this depends on the emotion type and is not consistent in all cases.

The two color measures have different effects. C_C shows its strongest patterns for feelings of boredom and safety, but its influence is weaker or depends on other factors for different outcomes. C_H, on the other hand, reflects how harmonious or tense the color palette feels, and its effects change in more complex ways across the four perception types. The random forest and partial dependence results reveal that these effects are not simply linear. Some outcomes exhibit sudden changes, U- or inverted-U-shaped patterns, and moderate “sweet spots” rather than straightforward trends. Also, C_C and C_H interact with each other. Some combinations strengthen the effect, while others reduce it. In Zurich, positive feelings are more likely in places with high harmony and low to moderate complexity. Negative feelings are stronger when complexity is high and harmony is low.

Comparing OLS and GWR models adds a spatial perspective. Citywide linear models capture only weak averages, whereas local models fit the data much better. These local models reveal that the link between color and emotion varies significantly across Zurich. The same color setup can create positive feelings in some areas but lead to clutter, monotony, or tension in others. A weak or non-significant citywide result does not mean there is no effect. Sometimes, it just means the average hides differences that change from place to place.

The Geneva study supports the overall validity of this approach, but only to a certain extent. The main patterns are similar in both cities. After controlling for scene content, the effect of color is still clear, especially for feelings of boredom and depression. However, the results in Geneva are not identical to those in Zurich. There are still differences depending on the outcome. For example, the feeling of safety shows little improvement, and the range of local effects is smaller. In general, the way color affects emotion seems to work in both cities, but its strength, exact shape, and how it varies by location depend on the context.

5.4 Limitations

1. **Model-based perception is not the same as resident survey ground truth.** The perception scores come from a trained prediction model rather than being

collected directly from Zurich or Geneva residents. This supports scalability, but it can introduce domain-shift and cultural/contextual bias (for example, Swiss streets compared with the contexts represented in the training data).

2. **Semantic controls are necessarily incomplete.** Semantic composition is controlled using proportional scene components (for example, vegetation, buildings, road, sky), but semantics can still “leak” through. Layout, fine-grained place cues, and activity intensity are not fully captured by coarse ratios.
3. **The cross-sectional design limits causal claims.** The analysis identifies associations and spatial heterogeneity, but it does not establish causality. Streets with certain color patterns may also differ systematically in unmodeled confounders—maintenance, socioeconomic context, noise, pedestrian flows, or land-use intensity—that could affect both color and perceived emotion.
4. **Spatial modeling choices introduce sensitivity.** Results from hotspot mapping, Moran’s I, and GWR depend on the spatial weights, bandwidth selection, and sampling design. While a connected KNN scheme and AICc-selected adaptive bandwidths are used here, alternative specifications could change the strength of detected clustering or the detailed pattern of local coefficients.

5.5 Future Work

1. **Add human validation and locally calibrate the perception model.** A focused perception survey in Zurich (and Geneva) could serve as a benchmark for the model outputs. It could also flag systematic faults and support local calibration, such as reweighting, bias correction, or fine-tuning for Swiss street contexts.
2. **Use stronger semantic “deconfounding” with finer scene representations.** Instead of mainly relying on ratio controls, subsequent research could use instance-level semantics, such as counts or areas of cars, pedestrians, or

storefronts. It might also include spatial configuration measures, such as facade continuity and enclosure proxies, as well as activity proxies, such as POI density, transit access, or nightlife indicators. This would make it easier to isolate what is truly color-specific.

3. **Expand color features and test other perceptual pathways.** Future analyses could add brightness and contrast statistics, warm–cool balance, saturation dispersion, palette entropy, and harmony subtypes to the existing C_C and C_H. This would create a clearer bridge to theories of aesthetic preference and affective response.
4. **Replicate temporally and seasonally.** Run the same pipeline over different seasons to test if the spatial patterns and net effects persist. This helps separate stable place effects from seasonal color shifts.
5. **Increase cross-city generalization with controlled comparisons.** The Geneva replication is a helpful first step. Extending to more Swiss cities with different urban morphologies would allow a more systematic test of what transfers across cities and what depends on local context.

6 Conclusion

This thesis set out to test whether the color of streetscapes sends a clear spatial signal in how people emotionally evaluate street environments. It used a workflow that can be scaled and repeated. Zurich was the main case study. The research combined large-scale Google Street View images with deep learning to infer perceptions (Safe, Beautiful, Boring, Depressing). Semantic segmentation was used to control for scene content. Two color features—color complexity (C_C) and color harmony (C_H)—were measured in CIELAB after standard processing. Three main findings emerged. First, emotional perceptions are not random. They form a structured 'perception geography,' with values clustering along the street network and creating clear hotspot and coldspot patterns. Second, streetscape color is also organized in space. C_C and C_H show different “color geographies” that do not simply map onto a good–bad scale. Complexity is more sensitive to variety and fragmented interfaces. Harmony reflects how well colors fit together and is more continuous in the city center and other visually coordinated areas. Third, color adds extra explanatory power beyond scene content in the models, showing a measurable effect of color itself. However, this effect depends on the outcome and location. Similar color patterns can lead to different emotional responses depending on the local street context. A smaller-scale replication in Geneva helped confirm the findings. Using the same model, the Geneva study showed similar results: a net effect of color after controlling for scene content and a mostly consistent pattern in the local results.

Overall, this thesis provides a method that others can use. It also provides evidence that urban color is more than just decoration. Color is a measurable part of how people experience streets, with spatial and context-dependent links to emotion. The results suggest that color planning should look beyond single colors. It should consider how richness and coherence are arranged across street networks. Future research should focus on local human validation and better control of scene content. It should also include seasonal and multi-city studies to improve understanding and generalizability.

Bibliography

Schneider, A., Friedl, M. A., & Potere, D. (2010). Mapping global urban areas using MODIS 500-m data: New methods and datasets based on “urban ecoregions”. *Remote Sensing of Environment, 114*(8), 1733–1746. https://doi.org/10.1016/j.rse.2010.03.003

Lu, Y., Sarkar, C., & Xiao, Y. (2018). The effect of street-level greenery on walking behavior: Evidence from Hong Kong. *Social Science & Medicine, 208*, 41–49. https://doi.org/10.1016/j.socscimed.2018.05.02

Beenackers, M. A., Kruize, H., Barsties, L., Acda, A., Bakker, I., Droomers, M., Kamphuis, C. B. M., Koomen, E., Nijkamp, J. E., Vaandrager, L., Völker, B., Luijben, G., & Ruijsbroek, A. (2024). Urban densification in the Netherlands and its impact on mental health: An expert-based causal loop diagram. *Health & Place, 87*, 103218. https://doi.org/10.1016/j.healthplace.2024.103218

Ma, X., Ma, C., Wu, C., Xi, Y., Yang, R., Peng, N., Zhang, C., & Ren, F. (2021). Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing. *Cities, 110*, 103086. https://doi.org/10.1016/j.cities.2020.103086

Van Den Bosch, M., & Ode Sang, Å. (2017). Urban natural environments as nature-based solutions for improved public health – A systematic review of reviews. *Environmental Research, 158*, 373–384. https://doi.org/10.1016/j.envres.2017.05.040

Brownson, R. C., Hoehner, C. M., Day, K., Forsyth, A., & Sallis, J. F. (2009). Measuring the built environment for physical activity: State of the science. *American Journal of Preventive Medicine, 36*(4 Suppl), S99–S123.e12. https://doi.org/10.1016/j.amepre.2009.01.005

Carmona, M., Gabrieli, T., Hickman, R., Laopoulou, T., & Livingstone, N. (2018). Street appeal: The value of street improvements. *Progress in Planning, 126*, 1–51. https://doi.org/10.1016/j.progress.2017.09.001

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. https://doi.org/10.1016/j.landurbplan.2018.09.015

Matsuoka, R. H., & Kaplan, R. (2008). People needs in the urban landscape: Analysis of *Landscape and Urban Planning* contributions. *Landscape and Urban Planning, 84*(1), 7–19. https://doi.org/10.1016/j.landurbplan.2007.09.009

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

Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design, 14*(1), 65–84. https://doi.org/10.1080/13574800802451155

Nasar, J. L. (1989). Perception, cognition, and evaluation of urban places. In I. Altman & E. H. Zube (Eds.), *Public places and spaces* (pp. 31–56). Springer. https://doi.org/10.1007/978-1-4684-5601-1_3

Alexander, C. (2004). *The nature of order: An essay on the art of building and the nature of the universe, book four: The luminous ground*. Center for Environmental Structure

Thurmann-Moe, D. (2017). *Kolorowa rewolucja: Architektura, wnętrza, moda*. Wydawnictwo Agora.

Elliot, A. J., & Maier, M. A. (2014). Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual Review of Psychology, 65*, 95–120. https://doi.org/10.1146/annurev-psych-010213-115035

Manav, B. (2017). Color-emotion associations, designing color schemes for urban environment-architectural settings. *Color Research and Application, 42*(5), 631–640. https://doi.org/10.1002/col.22123

Solli, M., & Lenz, R. (2011). Color emotions for multi-colored images. *Color Research and Application, 36*(3), 210–221. https://doi.org/10.1002/col.20604

Wilms, L., & Oberfeld, D. (2018). Color and emotion: Effects of hue, saturation, and brightness. *Psychological Research, 82*(5), 896–914. https://doi.org/10.1007/s00426-017-0880-8

Hu, K., Xu, Z., Wang, X., Wang, Y., Li, H., & Zhang, Y. (2023). Research on street color environment perception based on CEP-KASS framework. *Buildings, 13*(10), 2649. https://doi.org/10.3390/buildings13102649

Jiang, H., Lu, S., & Xiao, Y. (2022). Method of urban color evaluation for historic and cultural areas in Shanghai based on street view technology. *Urban Planning Forum*, (3), 111–118. https://doi.org/10.16361/j.upf.202203015

Chen, N., Xu, X., Tan, M., & Wang, X. (2022). A spatial analysis of urban color harmony in five global metropolises. *Journal of Resources and Ecology, 13*(2), 238–246. https://doi.org/10.5814/j.issn.1674-764x.2022.02.007

Zhai, Y., Gong, R., Huo, J., & Fan, B. (2023). Building facade color distribution, color harmony and diversity in relation to street functions: Using street view images and deep learning. *ISPRS International Journal of Geo-Information, 12*(6), 224. https://doi.org/10.3390/ijgi12060224

Subiza-Pérez, M., Vozmediano, L., & San Juan, C. (2020). Green and blue settings as providers of mental health ecosystem services: Comparing urban beaches and parks and building a predictive model of psychological restoration. *Landscape and Urban Planning, 204*, 103926. https://doi.org/10.1016/j.landurbplan.2020.103926

Duan, J., Wang, Y., Fan, C., Xia, B., & de Groot, R. (2018). Perception of urban environmental risks and the effects of urban green infrastructures (UGIs) on human well-being in four public green spaces of Guangzhou, China. *Environmental Management, 62*(3), 500–517. https://doi.org/10.1007/s00267-018-1068-8

Moore, G., Croxford, B., Adams, M., Refaee, M., Cox, T., & Sharples, S. (2006). Urban environmental quality: Perceptions and measures in three UK cities. In Ü. Mander, C. A. Brebbia, & E. Tiezzi (Eds.), *The sustainable city IV: Urban regeneration and sustainability* (*WIT Transactions on Ecology and the Environment*, Vol. 93, pp. 785–794). WIT Press. https://doi.org/10.2495/SC060751

Griew, P., Hillsdon, M., Foster, C., Coombes, E., Jones, A., & Wilkinson, P. (2013). Developing and testing a street audit tool using Google Street View to measure environmental supportiveness for physical activity. *International Journal of Behavioral Nutrition and Physical Activity, 10*, 103. https://doi.org/10.1186/1479-5868-10-103

Bonaiuto, M. (2004). Residential satisfaction and perceived urban quality. In C. D. Spielberger (Ed.), *Encyclopedia of applied psychology* (Vol. 3, pp. 267–272). Elsevier. https://doi.org/10.1016/B0-12-657410-3/00698-X

Liu, M., Han, L., Xiong, S., Qing, L., Ji, H., & Peng, Y. (2019). Large-scale street space quality evaluation based on deep learning over street view image. In Y. Zhao, N. Barnes, B. Chen, R. Westermann, X. Kong, & C. Lin (Eds.), *Image and graphics* (Lecture Notes in Computer Science, Vol. 11902, pp. 690–701). Springer. https://doi.org/10.1007/978-3-030-34110-7_58

Qiu, W., Li, W., Zhang, Z., Li, X., Liu, X., & Huang, X. (2021). *Subjective and objective measures of streetscape perceptions: Relationships with property value in Shanghai* (Version 1) [Preprint]. Preprints.org. https://doi.org/10.20944/preprints202103.0506.v1

He, J., Zhang, J., Yao, Y., & Li, X. (2023). Extracting human perceptions from street view images for better assessing urban renewal potential. *Cities, 134*, 104189. https://doi.org/10.1016/j.cities.2023.104189

Dai, M. (2021). Quantitative contrast of urban agglomeration colors based on image clustering algorithm: Case study of the Xia-Zhang-Quan metropolitan area. *Frontiers of Architectural Research, 10*(4), 692–700. https://doi.org/10.1016/j.foar.2021.07.006

Rui, Q., & Cheng, H. (2023). Quantifying the spatial quality of urban streets with open street view images: A case study of the main urban area of Fuzhou. *Ecological Indicators, 156*, 111204. https://doi.org/10.1016/j.ecolind.2023.111204

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In B. Leibe, J. Matas, & N. Sebe (Eds.), *Computer vision – ECCV 2016* (Lecture Notes in Computer Science, Vol. 9907, pp. 196–212). Springer. https://doi.org/10.1007/978-3-319-46466-4_12

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning, 180*, 148–160. https://doi.org/10.1016/j.landurbplan.2018.08.025

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*(12), 2363–2384. https://doi.org/10.1080/13658816.2019.1643024

Long, Y., & Ye, Y. (2019). Measuring human-scale urban form and its performance. *Landscape and Urban Planning, 191*, 103612. https://doi.org/10.1016/j.landurbplan.2019.103612

Smardon, R. C. (1988). Perception and aesthetics of the urban environment: Review of the role of vegetation. *Landscape and Urban Planning, 15*(1–2), 85–106. https://doi.org/10.1016/0169-2046(88)90018-7

Row, A. T., & Jacobs, J. (1962). *The death and life of great American cities*. *The Yale Law Journal, 71*(8), 1597–1604. https://doi.org/10.2307/794509

Law, S., Paige, B., & Russell, C. (2019). *Take a look around: Using Street View and satellite images to estimate house prices*. *ACM Transactions on Intelligent Systems and Technology, 10*(5), 1–19. https://doi.org/10.1145/3342240

Tran, T. (2024). *Real estate building detail enrichment with Street View image prediction* (Unpublished master's thesis). Lappeenranta-Lahti University of Technology, Finland.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). *Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States*. *Proceedings of the National Academy of Sciences, 114*(50), 13108–13113. https://doi.org/10.1073/pnas.1700035114

Yin, Y., Shao, Y., Xue, Z., Thwaites, K., & Zhang, K. (2020). *An explorative study on the identification and evaluation of restorative streetscape elements*. *Landscape Architecture Frontiers, 8*(4), 76–89. https://doi.org/10.15302/J-LAF-0-020005

Chen, L., Lu, Y., Sheng, Q., Ye, Y., Wang, R., & Liu, Y. (2020). *Estimating pedestrian volume using Street View images: A large-scale validation test*. *Computers, Environment and Urban Systems, 81*, 101481. https://doi.org/10.1016/j.compenvurbsys.2020.101481

Xie, H., Liu, L., & Yue, H. (2022). *Modeling the effect of streetscape environment on crime using Street View images and interpretable machine-learning technique*. *International Journal of Environmental Research and Public Health, 19*(21), 13833. https://doi.org/10.3390/ijerph192113833

Kadiyam, S. V. S. P. (2021). *Crime rate prediction from street view images using convolutional neural networks and transfer learning* (Master's thesis, University of Twente, Enschede, The Netherlands). Faculty of Geo-Information Science and Earth Observation.

Yue, H., Liu, L., Xu, C., Song, G., Chen, J., He, L., & Duan, L. (2024). *Investigating the diurnal effects of on-street population and streetscape physical environment on street theft crime: A machine learning and negative binomial regression approach using street view images*. *Applied Geography, 163*, 103194. https://doi.org/10.1016/j.apgeog.2023.103194

Najafizadeh, L., & Froehlich, J. E. (2018). *A feasibility study of using Google Street View and computer vision to track the evolution of urban accessibility*. In Proceedings of the 20th International ACM SIGACCESS Conference on

Computers and Accessibility (pp. 340–342). ACM. https://doi.org/10.1145/3234695.3240999

Weld, G., Jang, E., Li, A., Zeng, A., Heimerl, K., & Froehlich, J. E. (2019). *Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery*. In Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (pp. 196–209). ACM. https://doi.org/10.1145/3308561.3353798

Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). *Assessing street-level urban greenery using Google Street View and a modified green view index*. *Urban Forestry & Urban Greening, 14*(3), 675–685. https://doi.org/10.1016/j.ufug.2015.06.006

Li, X., Cai, B. Y., & Ratti, C. (2018). *Using street-level images and deep learning for urban landscape studies*. *Landscape Architecture Frontiers, 6*(2), 20–30. https://doi.org/10.15302/J-LAF-20180203

Li, X. (2021). *Examining the spatial distribution and temporal change of the green view index in New York City using Google Street View images and deep learning. Environment and Planning B: Urban Analytics and City Science, 48*(7), 2039–2054. https://doi.org/10.1177/2399808320962511

Zhang, W., & Zeng, H. (2024). *Spatial differentiation characteristics and influencing factors of the green view index in urban areas based on Street View images: A case study of Futian District, Shenzhen, China*. *Urban Forestry & Urban Greening, 93*, 128219. https://doi.org/10.1016/j.ufug.2024.128219

Nasar, J. L. (1994). Urban design aesthetics: The evaluative qualities of building exteriors. Environment and Behavior, 26(3), 377–401. https://doi.org/10.1177/001391659402600305

Weber, R., Schnier, J., & Jacobsen, T. (2008). *Aesthetics of streetscapes: Influence of fundamental properties on aesthetic judgments of urban space*. *Perceptual and Motor Skills, 106*(1), 128–146. https://doi.org/10.2466/pms.106.1.128-146

Quercia, D., O'Hare, N. K., & Cramer, H. (2014). *Aesthetic capital: What makes London look beautiful, quiet, and happy?* In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 945–955). ACM. https://doi.org/10.1145/2531602.2531613

Frank, L. D., & Engelke, P. O. (2001). *The built environment and human activity patterns: Exploring the impacts of urban form on public health*. *Journal of Planning Literature, 16*(2), 202–218. https://doi.org/10.1177/08854120122093339

Hong, J. Y., & Jeon, J. Y. (2015). *Influence of urban contexts on soundscape perceptions: A structural equation modeling approach*. *Landscape and Urban Planning, 141*, 78–87. https://doi.org/10.1016/j.landurbplan.2015.05.004

Ulrich, R. S. (1979). *Visual landscapes and psychological well-being*. *Landscape Research, 4*(1), 17–23. https://doi.org/10.1080/01426397908705892

Kaplan, R., & Kaplan, S. (1989). *The experience of nature: A psychological perspective*. Cambridge University Press.

Lynch, K. (1960). *The image of the city*. MIT Press.

Nasar, J. L. (1997). *The evaluative image of the city*. Sage Publications.

Tuan, Y.-F. (1977). *Space and place: The perspective of experience*. University of Minnesota Press.

Wei, J., Yue, W., Li, M., & Gao, J. (2022). *Mapping human perception of urban landscape from street-view images: A deep-learning approach*. *International Journal of Applied Earth Observation and Geoinformation, 112*, 102886. https://doi.org/10.1016/j.jag.2022.102886

Ji, H., Qing, L., Han, L., Wang, Z., Cheng, Y., & Peng, Y. (2021). *A new data-enabled intelligence framework for evaluating urban space perception*. *ISPRS International Journal of Geo-Information, 10*(6), 400. https://doi.org/10.3390/ijgi10060400

Wartmann, F. M., Frick, J., Kienast, F., & Hunziker, M. (2021). *Factors influencing visual landscape quality perceived by the public: Results from a national survey*. *Landscape and Urban Planning, 208*, 104024. https://doi.org/10.1016/j.landurbplan.2020.10402

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

Ordonez, V., & Berg, T. L. (2014). *Learning high-level judgments of urban perception*. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014: 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014, Proceedings, Part VI* (Vol. 8694, pp. 494–510). Springer International Publishing. https://doi.org/10.1007/978-3-319-10599-4_32

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. https://doi.org/10.1016/j.cities.2023.104704

Von Goethe, J. W. (1970). *Theory of colours* (C. L. Eastlake, Trans.; D. B. Judd, Intro.). MIT Press.

Lee, A. H. (2023). *What does colour tell about tourist experiences? Tourism Geographies, 25*(1), 136–157. https://doi.org/10.1080/14616688.2020.1852594

Zhang, J., Fukuda, T., & Yabuki, N. (2021). *Development of a city-scale approach for facade color measurement with building functional classification using deep learning and street view images*. *ISPRS International Journal of Geo-Information, 10*(8), 551. https://doi.org/10.3390/ijgi10080551

Zhong, T., Ye, C., Wang, Z., Tang, G., Zhang, W., & Ye, Y. (2021). *City-scale mapping of urban facade color using street-view imagery*. *Remote Sensing, 13*(8), 1591. https://doi.org/10.3390/rs13081591

Yang, R., Deng, X., Shi, H., Wang, Z., He, H., Xu, J., & Xiao, Y. (2024). *A novel approach for assessing color harmony of historical buildings via street view image*. *Frontiers of Architectural Research, 13*(4), 764–775. https://doi.org/10.1016/j.foar.2024.02.014

Boeri, C. (2017). *Color loci placemaking: The urban color between needs of continuity and renewal*. *Color Research & Application, 42*(5), 641–649. https://doi.org/10.1002/col.22128

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

Meng, Y., Sun, D., Lyu, M., Niu, J., & Fukuda, H. (2024). *Measuring human perception of residential built environment through street view image and deep learning*. *Environmental Research Communications, 6*(5), 055020. https://doi.org/10.1088/2515-7620/ad4e0e

Cheng, L., Chu, S., Zong, W., Li, S., Wu, J., & Li, M. (2017). *Use of Tencent street view imagery for visual perception of streets*. *ISPRS International Journal of Geo-Information, 6*(9), 265. https://doi.org/10.3390/ijgi6090265

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems, 65*, 126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PLOS ONE, 12*(8), e0180698. https://doi.org/10.1371/journal.pone.0180698

Moradi, M., Roche, S., & Mostafavi, M. A. (2022). Exploring five indicators for the quality of OpenStreetMap road networks: A case study of Québec, Canada. *Geomatica, 75*(4), 178–208. https://doi.org/10.1139/geomat-2021-0012

Chen, N., Wang, L., Xu, T., & Wang, M. (2025). Perception of urban street visual color environment based on the CEP-KASS framework. *Landscape and Urban Planning, 259*, 105359. https://doi.org/10.1016/j.landurbplan.2025.105359

Qi, Z., Li, J., Yang, X., & He, Z. (2025). How the characteristics of street color affect visitor emotional experience. *Computational Urban Science, 5*, 7. https://doi.org/10.1007/s43762-025-00167-z

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*(1), 94–100. https://doi.org/10.1016/j.amepre.2010.09.034

Wilson, J. S., Kelly, C. M., Schootman, M., Baker, E. A., Banerjee, A., Clennin, M., & Miller, D. K. (2012). *Assessing the built environment using omnidirectional imagery*. *American Journal of Preventive Medicine, 42*(2), 193–199. https://doi.org/10.1016/j.amepre.2011.09.029

Curtis, A., Locke, D. H., & Grogan, C. M. (2013). *Using Google Street View for systematic observation of the urban built environment*. *International Journal of Health Geographics, 12*, 53. https://doi.org/10.1186/1476-072X-12-53

Odgers, C. L., Donley, S. I., Caspi, A., Bates, C. J., Moffitt, T. E., & Arseneault, L. (2012). *Systematic social observation of children's neighborhoods using Google Street View: A reliable and cost-effective method*. *Journal of Child Psychology and Psychiatry, 53*(10), 1009–1017. https://doi.org/10.1111/j.1469-7610.2012.02565.x

Maslow, A. H. (1943). *A theory of human motivation*. *Psychological Review, 50*(4), 370–396. https://doi.org/10.1037/h0054346

Gong, Y., Palmer, S., Gallacher, J., Marsden, T., & Fone, D. (2016). *A systematic review of the relationship between objective measurements of the urban environment and psychological distress*. *Environment International, 96*, 48–57. https://doi.org/10.1016/j.envint.2016.08.019

Ito, K., Zhu, Y., Abdelrahman, M., Liang, X., Fan, Z., Hou, Y., Zhao, T., Ma, R., Fujiwara, K., Ouyang, J., Quintana, M., & Biljecki, F. (2025). *ZenSVI: An open-source software for the integrated acquisition, processing and analysis of street view imagery towards scalable urban science*. *Computers, Environment and Urban Systems, 119*, 102283. https://doi.org/10.1016/j.compenvurbsys.2025.102283

Ouyang, J. (2023). *Code repository for predicting human perception* [Computer software]. GitHub. Retrieved February 28, 2026, from https://github.com/strawmelon11/human-perception-place-pulse

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In International Conference on Learning Representations.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://papers.nips.cc/paper/7181-attention-is-all-you-need

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). *Scene parsing through ADE20K dataset*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5122–5130). IEEE. https://doi.org/10.1109/CVPR.2017.544

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022). *Masked-attention mask transformer for universal image segmentation*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1290–1299). IEEE. https://doi.org/10.1109/CVPR52688.2022.00135

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

Harada, K., & Oki, T. (2025). *How do weather and time of day affect street impression?* ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial

Information Sciences, X-4/W7-2025, 43–50. https://doi.org/10.5194/isprs-annals-X-4-W7-2025-43-2025

Ibrahim, M. R., Haworth, J., & Cheng, T. (2019). WeatherNet: Recognising weather and visual conditions from street-level images using deep residual learning. *ISPRS International Journal of Geo-Information, 8*(12), 549. https://doi.org/10.3390/ijgi8120549

Nguyen, L., Embrechts, J., & Teller, J. (2020). Colour in the city: A chromatic typology for the quantitative analysis of spatial coherence. *Landscape Research, 45*(1), 44–62. https://doi.org/10.1080/01426397.2019.1589437

Li, S., Chen, R., Yang, L., Huang, D., & Huang, S. (2020). Predictive modeling of consumer color preference: Using retail data and merchandise images. *Journal of Forecasting, 39*(8), 1305–1323. https://doi.org/10.1002/for.2689

Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis. In addition, AI tools such as ChatGPT, Gemini and Grammarly were used during the writing process to improve the clarity and readability of the text.

Nian Li

Nian Li

Zurich, March 26th, 2026