Impacts of Micro-Scale Built Environment Features on Residential Location Choice: a computer vision-aided assessment

CIE5060-09 MSc Thesis

Lanlan Yan

Impacts of Micro-Scale Built Environment Features on Residential Location Choice: a computer vision-aided assessment

by

Lanlan Yan-5456894

Thesis committee:

Chairperson	Dr.ir. Gonçalo Correia	TU Delft
Daily Supervisor	Dr.ir. Sander van Cranenburgh	TU Delft
Second Supervisor	Yiru Jiao	TU Delft

Cover Image: Taken by the Author

TU Delft

Preface

Studying Transport Planning at TU Delft has always been my dream for undergraduate studies. Reflecting on my Master's journey, the first year was stressful yet immensely rewarding, filled with new knowledge and transformative experiences. In the second year, I faced unexpected challenges during my graduation project, which forced me to reconsider my learning approach. Despite the tough process, it prompted me to think about my future and motivated me to pursue myself without being swayed by external influences.

Starting my graduation project, I was initially afraid of the challenges, especially given my lack of experience in Computer Vision and AI. However, I discovered great joy in the process, and the sense of accomplishment and happiness I felt at each stage far outweighed the difficulties I faced.

I am profoundly grateful to those who have supported me during this journey:

Firstly, I want to thank my thesis committee. My chairperson, Gonçalo Correia, thank you for your constructive advice during every meeting and for helping me improve my report writing. Since the operation research course, I have regarded you as an excellent professor. Attending your classes always feels rewarding but easy-to-follow at the same time. I am so delighted that you supervised both my bachelor's and master's theses. My daily supervisor, Sander van Cranenburgh, I appreciate that you not only pointed out issues that I could not realise, but also gave me positive feedback at each stage of achievement. Additionally, thank you for sharing valuable study materials that helped me quickly step into the world of computer vision and organising the graduation club to enable me to learn more about related and interesting projects. My second supervisor, Yiru Jiao, thank you for helping me to find the research scope that truly interested me after the termination of my last project. From bicycle route choice to residential location choice, you patiently guided me to the exciting field of computer vision. You gave me detailed feedback and encouraged me to believe in myself, which means a lot to me. Every discussion with you was enjoyable, and I admire your attitude as a researcher. In short, without your collective help, I could not have successfully completed this thesis.

I want to extend my deepest appreciation to my family, especially my beloved parents and puppy ”Mangguo,” for their unwavering support. Every time I feel discouraged, it is my parents' trust that gives me the confidence to keep going. Thank you for being my best friends. I am also incredibly grateful to my friends, who have been my constant companions throughout this journey.

Finally, I remind myself to embrace the future with courage, ready to face uncertainties and challenges head-on.

Lanlan Yan
Chongqing, June 2024

Summary

Studies have highlighted the impact of residential neighbourhood built environments (BE) on well-being in recent years since people spend significant time in these areas. These studies examine correlations between micro-scale BE features—such as trees, grass, fences, and bikes—and a single aspect of well-being (i.e., physical health, social interaction, mental health) or of perceptions (e.g., safety). However, they do not comprehensively explore how these features influence residential preferences nor provide clear guidance on designing micro-scale BE features to enhance the attractiveness of residential areas and overall well-being.

To build appealing residential neighbourhoods cost-efficiently and comprehensively, we should study people's preferences by analyzing residential location choice (RLC) datasets to understand how different micro-scale BE features influence attractiveness and well-being. However, few studies investigate the quantified impacts of these features on RLC.

van Cranenburgh and Garrido-Valenzuela (2023) conducted a stated choice experiment where respondents selected preferred residential neighbourhoods based on street view images (SVI). They introduced a framework, the Computer Vision-enriched Discrete Choice Model (CV-DCM). It uses a computer vision model to convert visual information into numerical data to integrate with traditional discrete choice models. However, the applied computer vision model is not for quantifying micro-scale BE features, limiting interpreting their impacts on RLC.

Building on van Cranenburgh and Garrido-Valenzuela (2023)'s framework and datasets (RLC and SVI datasets), this thesis proposes a semantic CV-DCM to study the effects of micro-scale BE features on RLC. The semantic computer vision model applied in this thesis is a panoptic segmentation model that incorporates instance and semantic segmentation for categories better quantified in units of instance and pixels, respectively. It can address the issue that some micro-scale BE features are challenging to count in images like trees.

After applying the semantic computer vision model, 400 generated masks are randomly selected and evaluated manually to examine whether the zero-shot computer vision model accurately quantifies micro-scale BE features in the new SVI dataset for subsequent choice modelling. Additionally, the detailed analysis of specific categories during mask evaluation provides valuable insights into the actual accuracies of different categories, enhancing the interpretation of estimated coefficients in choice modelling.

Overall, the choice modelling results highlight the specific impacts of various micro-scale BE features on RLC, providing valuable insights for urban planners. By pinpointing the most influential elements, the study facilitates the cost-effective restructuring of residential neighbourhoods to boost their attractiveness and enhance residents' well-being. For instance, restricting unattractive features like motorcycles in residential neighbourhoods and planting or maintaining more trees are the most attractive among vegetation.

Contents

List of Figures

List of Tables

1

Introduction

This chapter includes the background of this thesis, including the substantive and methodology gaps of previous relevant studies in section 1.1. Research objectives and questions are formulated in section 1.2. Section 1.3 outlines the structure of this report.

1.1. Background

In recent years, numerous studies have highlighted the potential impacts of a residential neighbourhood's built environment (BE) on people's well-being (Greene et al., 2020; Pfeiffer & Cloutier, 2016). Since individuals spend a significant amount of time in their neighbourhoods, these areas play a crucial role in providing spaces for physical activities and opportunities for social interactions (Ma et al., 2018; Pfeiffer & Cloutier, 2016; Sallis et al., 2022). Also, certain physical elements of BE correlate with residents' mental health (Y.-T. Wu et al., 2014). In general, previous research have investigated the correlations between various features of BE and different aspects of well-being (i.e., physical health, social interaction, and mental health). To be more specific, BE features are usually studied at the neighbourhood- or street level, where the former and latter indicate the macro-scale and micro-scale ones.

Macro-scale BE features encompass abstract concepts like the 5Ds (i.e., density, diversity, design, destination accessibility and distance to transit) proposed by Cervero et al. (2009) and Ewing and Cervero (2010). Various factors of the macro-scale BE features are also linked to specific facets of people's well-being. For instance, residential density, mixed land use, and street connectivity are highly correlated with walking frequency (Sallis et al., 2022). Streets with higher accessibility may attract runners, whereas higher job density might hinder running (Jiang et al., 2022).

Micro-scale BE features specifically indicate the physical elements that are visible at eye level in residential neighborhoods (e.g., surrounding trees, grass, fences, vehicles, bikes, road and sidewalks) (Sallis et al., 2022). Research demonstrates positive correlations between accessible green spaces (e.g., trees and grass) and activities like running (Jiang et al., 2022), cycling (Mertens et al., 2014), walking (Molina-García et al., 2020; Moniruzzaman & Páez, 2012; Steinmetz-Wood et al., 2019; 2020). Additionally, neighborhoods designed with traditional features like grid-lined streets, anterior garages, and front porches are positively correlated with social engagement among residents (Pfeiffer & Cloutier, 2016). On the contrary, factors such as litter-strewn streets, graffiti-ridden surfaces, and properties with broken windows create an atmosphere conducive to social disorder and crime (Perkins et al., 1992; Wilson & Kelling, 2017). A poor BE not only impedes daily activities but also amplifies stress and feelings of helplessness, posing significant risks to mental health (Blair et al., 2014; Y.-T. Wu et al., 2014).

Compared to modifying macro-scale features (Sallis et al., 2022; Steinmetz-Wood et al., 2020), adjusting micro-scale features of neighborhoods offers a cost-effective means of creating appealing residential environments. Micro-scale features are often more adaptable, as they do not necessitate a complete restructuring of the neighborhood layout, enabling swift application of findings to existing neighborhood settings (Cain et al., 2014).

However, studies on correlations between micro-scale BE features and well-being only focus on a particular aspect of well-being. These studies cannot offer comprehensive insights into how to design micro-scale BE features to enhance overall well-being (Jeon & Woo, 2023; Koo et al., 2022; Zhao et al., 2023). Moreover, these studies merely explore associations between micro-scale BE features and specific well-being aspects rather than examining their direct impacts on people's behavior. The absence of causality in these findings complicates drawing conclusions about whether certain features positively or negatively influence people's behavior.

To build appealing residential neighborhoods in a cost-efficient way that account for all aspects of well-being, policymakers and urban designers should obtain valuable insights into the specific types, quantities, and combinations of micro-scale BE features that enhance overall attractiveness. By investigating how these features influence individuals' choices regarding where to live, we can gain a comprehensive understanding of the quantified impacts of all micro-scale BE features on people's preferences for residential neighborhoods. Therefore, it's essential to study people's residential preferences by analyzing datasets of residential location choices (RLC) (Ma et al., 2018).

Under the context of BE and RLC, most studies investigate how macro-scale BE features influence RLC, like land use mixture, distance to the city centre, transit and other facilities (Diao et al., 2016; Guan & Wang, 2020; Kim, Boxall, et al., 2019; K. Wang & Ozbilen, 2020). Nevertheless, micro-scale BE features often receive inadequate attention. Three substantive gaps emerge in the few studies on the influence of micro-scale BE features on RLC. Firstly, studies often concentrate on specific elements, neglecting others in open spaces, hindering a comprehensive comparison of their effects on RLC and understanding of preferences (Cockx & Canters, 2020). Secondly, the categorization of micro-scale BE features in RLC modelling studies remains inadequate (e.g., vegetation can be further categorized into trees, grass, etc.), obscuring impacts on choice behaviour. Lastly, scholars tend to examine the presence of the physical elements rather than quantifying them. They rely on subjective assessments, posing challenges in objective measurement, particularly for amorphous elements like trees and buildings that are uncountable (Cockx & Canters, 2020). Notably, the three limitations are attributed to the methodology gap of quantifying the micro-scale BE features. It's time-consuming for humans to identify all micro-scale BE features in images whilst quantify them since there are some uncountable features. They can only pay attention to certain features and examine whether they are present. These limitations underscore the need for more holistic and quantifiable approaches to studying the impact of the micro-scale BE features on RLC.

The methodology gap of quantifying micro-scale BE features also exist in early studies on associations between micro-scale BE features and daily activities or social interaction (Molina-García et al., 2020; Steinmetz-Wood et al., 2019; 2020). Scholars usually examine micro-scale BE features by virtual tools or field visits, only recording the presence of features instead of quantifying them. In later studies, with the development of new technologies and data sources, researchers apply the the combination of street view image data and semantic computer vision technology, to address the methodology gap. Street-view images (also known as street-level images) serve as a valuable and publicly accessible data source, offering a unique opportunity to examine visual features from a human perspective horizontally, a viewpoint not readily available in other common data sources like aerial or satellite imagery (Biljecki & Ito, 2021). Semantic computer vision models usually involve semantic segmentation, object detection and instance segmentation tasks, which can measure the objects in images with units of proportions of pixels or/and instances. Semantic computer vision models are used to quantify micro-scale BE features in street view images, studying the relationships between the features and different daily activities including running (Jiang et al., 2022), walking (Koo et al., 2022), cycling (Zhao et al., 2023). The emerging combination can also be adopted to study the impacts of micro-scale BE features on RLC more deeply and broadly.

The emerging combination of SVI and semantic computer vision models has been applied in studies on understanding people's perceptions of micro-scale BE features (Meng et al., 2024; Ramírez et al., 2021; Rossetti et al., 2019; Z. Wang et al., 2024; F. Zhang et al., 2018). These studies commonly use a large-scale SVI dataset: Place pulse 2.0, which consist of 110,988 images from 56 cities and 1,170,000 pairwise comparisons answered by 81,630 online survey participants on six perception scores. However, while these studies offer valuable insights, they often lack clear guidance for designing appealing BE near residences or identifying the most influential elements. Various perceptions shape preferences for residences, each perception carrying different weights in decision-making. For instance, although an open environment may seem inviting initially, safety concerns might lead individuals to prioritize safety over openness. Additionally, conflicting impacts of certain physical elements on different perceptions add complexity to understanding people's preferences. Therefore, generating policy-relevant information on BE design necessitates a focus on studying the impacts of physical elements on RLC rather than solely relying on people's perceptions.

To investigate the impacts of micro-scale BE features on RLC, a designated dataset capturing participants' housing preferences is essential. In a recent study by van Cranenburgh and Garrido-Valenzuela (2023), a novel Stated Choice Experiment was conducted, where participants made selections among various residential options, considering trade-offs between commute time (numeric attributes), monthly housing cost (numeric attributes) and street-level conditions shown in street view images (SVI). Additionally, they introduced a pioneering framework called the Computer Vision-enriched Discrete Choice Model (CV-DCM). This framework incorporates computer vision techniques to translate visual information into numerical data, which is then integrated into traditional discrete choice models alongside data from stated preference surveys.

However, it's noteworthy that their computer vision model was not designed to identify and quantify micro-scale BE features in images. Instead, it employed a feature extractor, a deep neural network trained to derive relevant features from images. This computer vision model generated the feature map (a.k.a. embedding) representing salient image features, which is a flat array of floating points. However, the interpretation of these elements of feature maps in terms of behavioral significance remained ambiguous. Consequently, the study can only tell the visual attractivesness (i.e., utilities) of images of residential neighborhoods rather than providing insights into actual choice behaviors.

1.2. Research objective and questions

Understanding how micro-scale BE features influence RLC and individuals' trade-offs among these features is crucial for informing better BE designs in residential areas and enhancing overall resident well-being. To address the gap of the work of van Cranenburgh and Garrido-Valenzuela (2023), this thesis proposes a Semantic CV-DCM, which can take semantic texts as prompts to identify and quantify micro-scale BE features. Building upon the dataset provided by van Cranenburgh and Garrido-Valenzuela (2023) and following their established framework, the Semantic CV-DCM comprises two primary components: firstly, identifying and quantifying micro-scale BE features using semantic computer vision models, and secondly, integrating the results of image parametrization into discrete choice models to explore people's preferences more deeply.

The proposed Semantic CV-DCM endeavors to furnish policymakers and urban planners with comprehensive insights into the design of all pertinent micro-scale BE features within residential areas, aimed at enhancing the overall appeal of these locales. Furthermore, beyond RLC, the Semantic CV-DCM has the potential to extend to other areas of choice modelling involving image parametrization, thereby contributing to advancements across various fields.

In order to achieve the research objective, the main research question of the thesis is presented below, which will be further explained in the following paragraphs:

Main Research Question:
What are the impacts of the micro-scale built environment features on the residential location choice?

Residential location choice (RLC) refers to how individuals or households decide where to live, considering factors such as affordability, proximity to work, access to amenities, safety, community atmosphere, and the surrounding environment (Cockx & Canters, 2020). While some factors like housing costs and proximity to work are easily measured, others, such as community atmosphere and the built environment, are more challenging to quantify. In this thesis, the focus is on utilizing image data to manifest the built environment (BE) features, which encompass physical elements commonly found in residential streets, including trees, houses, cars, street lights, plants, and bicycles. Through the analysis of these physical elements, this research aims to uncover their influences on RLC, providing valuable insights for decision-makers and stakeholders involved in urban planning and residential development.

Sub-questions:
To better answer the main research question, three parts are investigated in three sub-questions, individually, as follows:

1. **What micro-scale BE features are quantified by the semantic computer vision model?**

2. **To which extent can the semantic computer vision model accurately quantify micro-scale BE features?**

3. **How do micro-scale BE features and other attributes of residences affect people's choice behavior on residential locations?**

The first sub-question addresses the determination of texts to be provided to the semantic computer vision model. As mentioned earlier, this thesis employs a semantic computer vision model to identify and label micro-scale BE features in images. Consequently, it is essential to define micro-scale BE features in advance and provide corresponding semantic texts to the semantic computer vision model. A relevant study that investigated the effects of micro-scale BE features on perceptions of urban spaces has predefined eleven semantic texts, including ”building,” ”sidewalk,” and ”vegetation” (Rossetti et al., 2019). While some physical elements of the BE, such as trees, may be acknowledged to impact RLC by enhancing nearby residences' attractiveness, the influence of most elements on choice behaviour remains uncertain and requires empirical validation. Thus, there is a need to include as many physical elements of the BE as possible to account for potential influences that may not be initially apparent. However, it may be unnecessary to include specific physical elements if the number of images containing them is insufficient.

Choosing relevant micro-scale BE features of RLC is essential. One object may contain other objects. For instance, buildings are comprised of windows, walls and doors. It is questionable to choose which among ”window”, ”wall”, ”door”, and ”building” to represent the object ”building”. Windows and walls are considered opportunities for natural surveillance from residents and barriers that can divide space (e.g., private and public spaces) to promote territoriality, respectively (Zhanjun et al., 2022). However, as semantic computer vision models can only assign one category to each pixel/region, providing redundant categories may confuse the model and decrease the model's accuracy.

The second sub-question focuses on assessing the results of the semantic computer vision model applying to the Street View Image (SVI) data depicting residential areas in the Netherlands. Specifically, this evaluation aims to gauge the model's proficiency in accurately quantifying micro-scale Built Environment (BE) features depicted in these images.

Notably, the SVI dataset sourced from van Cranenburgh and Garrido-Valenzuela (2023) lacks manually-labeled pixels with specific semantic meanings, making direct training of the chosen semantic computer vision model using this dataset unfeasible. As a result, the chosen approach involves employing a zero-shot semantic computer vision model for this task. Unlike conventional models trained on predefined categories, a zero-shot model is trained to recognize objects or perform tasks without explicit examples or training data for those specific objects or tasks. A zero-shot model is beneficial when collecting labelled training data for all possible classes or tasks that are impractical or expensive (Cao et al., 2020). It allows models to generalize knowledge and adapt to new classes or tasks without additional training data.

However, this approach carries inherent risks. One critical consideration is whether the zero-shot semantic computer vision model can accurately quantify micro-scale BE features from images it has yet to be explicitly trained. Addressing this concern is crucial, as it directly influences the reliability of the quantified results in informing subsequent discrete choice models. Manually evaluating the results of some randomly chosen images can provide direct evidence to affirm the results of the semantic computer vision model, which can be input for choice modelling. Also, having a picture of how the semantic computer vision model performs on this SVI data may help interpret the discrete choice model's estimates if some estimates are unreasonable.

The third sub-question pertains to the analysis of results from model estimation, which is pivotal in this thesis as it delves into people's preferences for different built environment (BE) features when selecting residential locations. Answers to this sub-question offer insights into whether certain attributes extracted from images, such as the presence of trees or cars, positively or negatively affect respondents' decision-making processes. By examining and comparing the impact of various attributes on respondents' behavior, we can discern which attributes wield the greatest and least influence on residential location choices, as well as the elasticities (i.e., trade-offs) between them. Furthermore, this sub-question enables a comparison of the effects of micro-scale BE features and other attributes (e.g., commuting time and housing cost) on residential location choice, shedding light on how individuals balance considerations such as surrounding BE, location, and housing prices when making decisions.

1.3. Thesis outline

The rest of the thesis is organised as follows. Section 2 will review literature that are relevant with residential location choice, image use in discrete choice modelling, image parametrization of public spaces, panoptic segmentation. Section 3 will introduce methodology workflow, including the specific methods for the semantic computer vision model and discrete choice model, and how they connect with each other. Section 4 will present the results of semantic computer vision models and discrete choice models with corresponding analyses. Section 5 will discuss other findings. Finally, the last section will offer the conclusion.

2

Literature Review

This chapter presents a literature review within the research scope by discussing several important topics: (1) the introduction to residential location choice (RLC) including factors of built environment influencing RLC; (2) comprehensive review of macro-and micro-scale built environment features that are correlated to different aspects of people's well-being; (3) existing methods for quantify micro-scale BE features; (4) studies that use semantic computer vision (CV) models and tasks for quantification; (5) image use in discrete choice modelling; (6) other attributes of images (low-level features and season variations) that might affect choices.

2.1. Residential location choice

Residential location choice (RLC) refers to the decision-making process individuals or households undergo when selecting a place to live. Identifying a relevant set of household and location variables is necessary when modelling residential location choice (Cockx & Canters, 2020). Hurtubia et al. (2010) and Schirmer et al. (2014) provide a comprehensive review of related attributes in RLC modelling. Household variables typically include the age of the household head, household size, the presence of children, education level and the number of workers within the household. These attributes are essential in RLC modelling since similar households regarding income class or ethnic groups usually have similar RLCs and gather together. Location variables usually include the type of housing unit (e.g., attached or single-family houses with monthly housing costs), the neighbourhood conditions (e.g., land use mixture, population and job densities), and transport and access characteristics (e.g., distance to transit stations, commuting time). Neighbourhood conditions and transport and access characteristics are also usually denoted as the built environment (BE) in the context of residential location choice (RLC). Household variables will not be included in this thesis.

In most previous studies on the impacts of BE on RLC, BE is a wider description, encompassing factors such as the crime and traffic congestion (Ayoola et al., 2023), land use mixture (K. Wang & Ozbilen, 2020), home-centre distances (Guan & Wang, 2020) and so on. There are other studies defining BE as an "enormous"physical element located relatively far from the residence. The elements usually include parks (Czembrowski & Kronenberg, 2016), landmark buildings (Been et al., 2016), transportation infrastructures (e.g., airports and railways) (Diao et al., 2016; Yassin et al., 2019), and so on. However, very few studies have investigated “small”physical elements of BE that are very close to the residence.

The study of Cockx and Canters (2020) narrows the size of physical/visual elements. They introduce small elements related to the environment (e.g., covering the appearance of buildings, tidiness, air quality and quietness) and facilities (e.g., covering streets, pavements, cycle paths, green, public transport, stores, health care, social services, daycare, and culture and recreation) and incorporate them as variables into the residential location choice modelling. However, variables describing the environment and facilities are ordinal variables with a natural ranking: “unpleasant”, “satisfactory”, and “very pleasant”for the environment; “very bad condition”, “very normal condition”, and “very excellent condition”for the facilities. The environment and facilities are evaluated based on subjective opinions rather than objective statistics/numbers.

A key finding from the aforementioned studies is the inadequacy of quantifying the ”small” physical elements of the BE in existing research on RLC. Typically, scholars focus on quantifying access or proximity to ”enormous” physical elements, such as distances to parks (Czembrowski & Kronenberg, 2016) and railways (Diao et al., 2016). Additionally, some studies utilize binary variables to represent the presence of specific attributes, such as houses with waterfronts (Kim, Boxall, et al., 2019). Similarly, studies like Cockx and Canters (2020) fail to objectively quantify ”small” physical elements of BE, which are evaluated subjectively. Learning how previous studies quantify BE features in the residential location choices is essential as most of them implement quantifying by pre-defining an order for the attribute and assigning a value to the BE feature based on the order.

2.2. Macro- and micro-scale built environment features

Built environment features are typically categorized into two scales: macro- and micro-scale. Macro-scale features encompass neighbourhood-level conditions, commonly referred to as the 5Ds: density, diversity, design, destination accessibility, and distance to transit. As noted in the preceding subsection, these factors are frequently studied in research on residential location choice as influential factors. Additionally, various studies have identified significant correlations between macro-scale built environment features and activities such as walking (Koo et al., 2022), running (Jiang et al., 2022), overall health (L. Zhang et al., 2023).

Micro-scale features pertain to street-level conditions, encompassing the physical elements and features that people can directly perceive on the streets. Research often explores these features in the context of various aspects of daily life, including health, street vitality, crime rates, and air quality (Molina-García et al., 2020; Qi et al., 2022; Steinmetz-Wood et al., 2019; 2020; W. Wu et al., 2022; Y.-T. Wu et al., 2014; Zhanjun et al., 2022). Such studies consistently reveal significant correlations between micro-scale built environment features and these contextual factors. However, despite their relevance to residential location choice, micro-scale features—”small” physical elements of the built environment—are seldom investigated in this context, as highlighted in the preceding subsection.

Nevertheless, a neighbourhood’s support for daily activities like walking, running, cycling, street vitality, and a safe atmosphere are crucial considerations in residential decision-making.

The systematic impact of micro-scale built environment features on residential location choice remains largely unexplored. Given that existing studies on micro-scale features primarily focus on daily activities, overall health and crime rates, the upcoming literature review will concentrate on these studies to explore how they extract the qualities of micro-scale built environment features depicted in images.

2.3. Existing methods for quantifying micro-scale BE features

In previous studies, micro-scale built environment features have predominantly been assessed through field investigations, where scholars gather information directly from the real-world environment rather than relying solely on existing records or secondary sources (Adkins et al., 2012; Loukaitou-Sideris et al., 2001). This approach entails physically going out into the field, whether a geographical location, a specific site, or an area of study, to observe and document phenomena firsthand, incurring high labour and time costs.

To mitigate these challenges, some studies have explored alternative measurement approaches, leveraging virtual audit tools such as MAPS (Sallis et al., 2022), Virtual-STEPS (Steinmetz-Wood et al., 2020), and the Residential Environment Assessment Tool (Y.-T. Wu et al., 2014). These initiatives utilize Street View Imagery (SVI) data, providing a comprehensive view of the physical elements in the built environment (BE), offering a close approximation of residents' surroundings (Li et al., 2022). Researchers conduct detailed audits of micro-scale BE features in neighbourhoods, assessing images based on their suitability for various behaviours like walking or cycling, as well as mental health indicators (Mertens et al., 2014; Steinmetz-Wood et al., 2020; Y.-T. Wu et al., 2014).

However, the above both methods rely on human assessment, making them subjective, time-consuming, and labour-intensive. Furthermore, these approaches often focus on recording the presence of features rather than quantifying them objectively or including all eye-level physical elements. Table 2.1 shows studies that use virtual tools to quantify micro-scale BE features, their studying contexts and some of including features as examples.

A promising advancement emerges in the form of semantic computer vision models based on deep learning, enabling the objective quantification of micro-scale BE features in an automated manner. This method identifies the presence of specific environmental elements and discerns their proportions, such as whether the green plants are lush. SVI data encompass a rich array of ground elements within extensive regions, such as trees, sky, buildings, roads, and grass, facilitating rapid data collection. Combining SVI data with semantic computer vision models allows researchers to overcome the limitations of traditional methods, enabling more refined, efficient, and human-centric perception research (Xu et al., 2023). Leveraging this emerging integration, researchers have already delved deeper into the relationships between micro-scale BE features and people's behaviour and health, yielding more robust findings. Tables 2.2 and 2.3 demonstrate recent papers that use semantic computer vision models and SVI to quantify micro-scale BE features.

Table 2.1: Relevant studies that use virtual tools to quantify micro-scale BE features

Literature	Contexts	Micro-scale BE features
Sallis et al. (2022)	all physical activities	Presence of transit stops, driveways, dustbins, benches, bicycle racks, traffic calming (roundabouts and speed bumps), crosswalk amenities (e.g., marked crosswalk), curb ramps, crossing signals, high and low streetlights, sidewalk...
Steinmetz-Wood et al. (2020)	walking	Presence of traffic calming (e.g., stop signs), aesthetics/disorder (e.g., graffiti), natural sights, signs of disorder, litter. Building height setback, building height road width ratio, building aesthetic design…
Molina-García et al. (2020)	walking	Average number of traffic lanes; number of regulated crossings; average parking street buffer, number of traffic calming features, aesthetic and social characteristics (e.g., building maintenance, fountains, sculptures, or art)...
Steinmetz-Wood et al. (2019)	walking	Pedestrian infrastructure (e.g., sidewalk, pedestrian crossing sign), traffic calming (e.g., presence of traffic lights, number of parking lanes), building characteristics (e.g., building setback and height), transit (presence of transit, type of transit), bicycling infrastructure (e.g., bike lanes and buffer), aesthetics/disorder (e.g., trees, litter, graffiti)
Mertens et al. (2014)	cycling	Traffic level (presence of driving cars), traffic calming (presence of a speed bump), evenness of the cycle path (good/poor), general upkeep (overall maintenance degree, e.g., graffiti, broken windows), vegetation (presence of trees along the road and greenery on houses), separation between cycle path and motorized traffic/sidewalk, and width of the cycle path (narrow or wide)
Y.-T. Wu et al. (2014)	mental health	Property level (e.g.,vacant properties, abandoned cars, low property maintenance, no trees in front gardens), street level (e.g., illegal parking, dog litter, poor path condition, no neighbourhood watch signs, public parking, front outlook: green/commercial/industrial)
Moniruzzaman and Páez (2012)	walking	Low or high (postive or negative) uses insegment, slope, sidewalk completeness, sidewalk connectivity, amenities, way findingaids, trees shading walking area, overall cleanliness/building maintenance, building height….

When comparing the micro-scale built environment (BE) features presented in Table 2.1 and Table 2.2 (or Tabel 2.3), a clear distinction emerges. While quantifying with virtual tools by humans allows for the discernment of the mere presence or absence of features and subjective assessments of their relative prominence, utilizing semantic computer vision models offers a more accurate quantization. These models not only identify specific objects but also measure the sizes of features within images, providing precise measurements in terms of pixel proportions. However, it is essential to acknowledge that quantifying features with virtual tools captures additional details such as building height, setback distances, and the presence of marked crosswalks. These nuanced qualities of the BE may pose challenges for semantic computer vision models to identify accurately. Nonetheless, a virtual tool often requires significant time for image processing and evaluation by human operators.

Table 2.2: Relevant studies that use semantic computer vision models and SVI to quantify micro-scale BE features

Literature	Contexts	Semantic Computer Vision Models	Micro BE features	Analysis method
Chen et al. (2024)	urban green space	pre-trained semantic segmentation model	sidewalk, person, street light, grass, tree, chair, sky, water, fence, wall, earth, fountain, mountain, sculpture, bridge	spatial regression models (ordinary least squares (OLS))
Zhao et al. (2023)	cycling volume	pre-trained Mask_RCNN (object detection) and Pyramid Scene Parsing Network (PSPNet, semantic segmentation)	tree, road, grass, car, streetlight, wall, building, sidewalk, earth, water, plant, awning, van, person, bridge, railing, bicycle, minibike, ceiling, chair	OLS regression
Fan et al. (2023)	health, crime, transport, and poverty	Pre-trained semantic segmentation model	street furniture, sidewalk, facade, window & opening, road, sky, grass and shrubs, trees, people, bike, vehicles (each combined several original categories in ADE 20 dataset)	least absolute shrinkage and selection operator (LASSO) regression
L. Zhang et al. (2023)	covid 19	DeepLabV3+ (semantic segmentation)	sky, building, road, wall, macrophanerophytes, bush, grass	linear regression
Jiang et al. (2022)	running	pre-trained PSPNet (semantic segmentation)	Wall, building, tree, road, grass, sidewalk, earth, plant, car, fence, signboard, awning, streetlight, van, ashcan, railing, person, minibike, chair, sculpture, bicycle, column, bridge, water, fountain, windowpane, mountain, ceiling, booth, sofa, lamp, skyscraper, lake, bulletin board, desk, pier, Sky view factor (SVF)	OLS regression

Table 2.2 – continued from previous page

Literature	Contexts	Semantic Computer Vision Models	Micro BE features	Analysis method
Yue et al. (2022)	health (chronic diseases and mental health)	Convolutional Neural Networks (ConvNets, scene label)	street greenness (trees and landscaping comprised at least 30% of the image), (2) presence of a crosswalk, (3) single lane road, (4) building type (single-family detached house vs. other), and (5) visible utility wires	linear regression
Zhanjun et al. (2022)	crime rates	Pre-trained fully convolutional network (FCN, semantic segmentation)	wall, fence, window, streetlight, building, street furniture, greenness, tree, car, person, ashcan, signboard, bench, pavement, road, building	regression models, GWR and MGWR,
Qi et al. (2022)	NO_2	Pre-trained PSPNet (semantic segmentation)	built environment, transport network, transport vehicles, natural, vegetation, water, and human (each one combined several relevant categoires)	land Use regression model
Nguyen et al. (2022)	patient health	ConvNets (scene label)	building type (the presence of any non-single-family detached house: yes/no), roads with a single lane (yes/no), crosswalk presence (yes/no), street greenness (at least 30% of the image consisted of trees and landscaping: yes/no), and the presence of visible utility wires overhead (yes/no)	poisson regression models
Koo et al. (2022)	walking	Pre-trained PSPNet (semantic segmentation)	building-to street-ratio (the ratio of the proportion of buildings and houses to the sum of the proportion of sidewalk, road, and car). The greenness (the sum of the proportion of tree, grass, and plant). Sidewalk-to-street proportion (the proportion of the share of sidewalk to the sum of the share of sidewalk, road, and car)	logistic regression models

Table 2.2 – continued from previous page

Literature	Contexts	Semantic Computer Vision Models	Micro BE features	Analysis method
Jeon and Woo (2023)	walking	Pre-trained HRNetV2-W48 model (semantic segmentation)	visual greenery ((tree + grass + plant)/total), outdoor openness (sky/total), street pavement (sidewalk/(road + sidewalk))	logistic regression models
H. Zhou et al. (2021)	drug places	Pre-trained PSPNet (semantic segmentation)	building, terrain, traffic sign, traffic light, pole, road, sidewalk	logistic regression models

Table 2.3: Studies quantifying micro-scale BE features by semantic computer vision models and Place Pulse 2.0 by Dubey et al. (2016)

Literature	Computer vision models	Micro BE beatures
Rossetti et al. (2019)	Pre-trained SegNet method (semantic segmentation)	building, car, cyclist, fence, pedestrian, pole, road, sidewalk, sky, traffic sign, and vegetation
Ramírez et al. (2021)	Pre-trained SegNet (semantic segmentation) and faster R-CNN (object detection)	semantic segmention (cyclist, building, car, fence, sidewalk, pedestrian, pole, road, traffic sign, sky and tree), object detection (car, person, truck, potted plant, bus, train, motorcycle, bicycle, traffic light, bench stop, sign, fire hydrant, umbrella, chair)
Meng et al. (2024)	Pre-trained DeeplabV3+ (semantic segmentation)	vehicle occurrence rate, enclosure, greenness, pedestrian occurrence rate, openness, natural landscape, Natural to artificial ratio of the vertical interface, Natural to artificial ratio of the horizontal interface (each one combined several relevant categories)
Z. Wang et al. (2024)	ResNet101	
F. Zhang et al. (2018)	Pre-trained ResNet50 (semantic segmentation)	wall, building, sky, tree, road, grass, sidewalk, plant, car, sign, stairs, van

2.4. Quantify micro-scale BE features by semantic CV model and SVI data

The studies listed in Table 2.2 predominantly explore the correlations between micro-scale built environment (BE) features and various aspects of human behaviour, health, and crime using regression models. However, while these studies shed light on correlations, they often need to uncover the direct influence of these features on people's physical or mental well-being.

For example, in the research conducted by Chen et al. (2024), which examines the hot spots of urban green spaces (UGS) and BE features, it is evident that proximity plays a pivotal role in people's decisions to visit green spaces. The convenience of accessing green spaces near residential, work, or leisure areas significantly influences individuals' choices regarding outdoor leisure activities. Overlooking the importance of proximity in studying UGS popularity and usage patterns can lead to an incomplete understanding of their appeal and utilization dynamics.

Incorporating proximity data into analyses provides a more holistic view of how and why people utilize different green spaces within urban environments. This comprehensive understanding can inform urban planning efforts, ensuring that green spaces are strategically located to cater to the needs of local communities. While spatial regression models utilized in this study aim to unveil statistical associations between variables, considering spatial dependencies, they do not inherently establish causality. On the contrary, choice modelling focuses on understanding causal relationships by examining how variations in independent variables impact choices or decisions. Typically, choice modelling employs experimental or quasi-experimental designs to establish causality between variables.

The Place Pulse 2.0 dataset (Dubey et al., 2016) presents an invaluable opportunity for developing deep learning models to predict urban perceptions based on visual features of the environment. Comprising over 1.17 million pairwise comparisons across approximately 110,988 images from cities worldwide, this dataset offers insights into perceptions across six key dimensions: safety, liveliness, boredom, wealth, depression, and beauty (Dubey et al., 2016). By replacing traditional low-throughput survey methods like questionnaires, the dataset enables rapid and large-scale estimation of residents' true perceptions regarding the environmental quality of their neighbourhoods (Z. Wang et al., 2024). Its extensive coverage and comprehensive perceptual dimensions make it a valuable resource for researchers seeking to understand the impacts of the built environment on perceptions, which has been leveraged in many studies. Table 2.3 demonstrates recent studies using the Place Pluse 2.0 dataset.

2.5. Semantic computer vision models and tasks for quantification

Most studies in Table 2.2 and 2.3 use pre-trained semantic segmentation models. For instance, in the work of Rossetti et al. (2019), a novel approach is proposed to quantify landscape-perception relations through discrete choice models. Semantic segmentation of images of public spaces, generated via machine learning algorithms, serves as the primary explanatory variable. These models, estimated using the Place Pulse dataset, offer insights into how users perceive the built environment based on its features. Noteworthy is how Rossetti et al. (2019) parametrizes images and extracts interpretable features, categorizing them into low-level (e.g., edges, number of binary large objects) and high-level ones (i.e., BE features).

Ramírez et al. (2021) integrates object detection and heterogeneity into choice modelling, expanding on the work of Rossetti et al. (2019). This study employs three types of features: colour and edge statistics, semantic segmentation, and object detection, each capturing distinct information. This study conducts three discrete choice model estimations: an initial model akin to the study of Rossetti et al. (2019), a second model incorporating bounding boxes of relevant objects for correcting segmentation errors, and a third model considering heterogeneity by constructing additional variables.

Building upon the study of Ramírez et al. (2021), which utilized semantic segmentation to quantify BE features, Rossetti et al. (2019) incorporates object detection into the semantic computer vision model for correcting segmentation errors. However, a limitation in the study of Rossetti et al. (2019) is that several BE features are labelled twice by two semantic computer vision tasks (i.e., semantic segmentation and object detection) since they did not assign various BE features to suitable semantic computer vision tasks. BE features should be categorized well to enhance fusion between semantic segmentation and object detection, and appropriate extraction processes should be selected based on quantification units. Semantic segmentation suits pixel-level category assignments, while object detection or instance segmentation is ideal for categories described by instance counts, such as cars and trains. Categories that are challenging to quantify by instance count, like trees, are better suited to semantic segmentation.

The above paragraph introduced two essential semantic computer vision tasks: semantic segmentation and object detection, which can quantify given image categories. Semantic segmentation captures the proportions of pixels labelled with corresponding categories, while object detection provides the counts of categories in an image (Ramírez et al., 2021). Another semantic computer vision task, instance segmentation, combines the principles of semantic segmentation and object detection to determine the number of instances and the pixel count for each instance for each category (Kirillov et al., 2019). It is evident that while object detection and instance segmentation excel in quantifying countable objects, they are less suitable for amorphous entities, unlike semantic segmentation, which is applicable across all categories.

The concept of panoptic segmentation, as introduced by Kirillov et al. (2019), integrates both instance and semantic segmentation to offer a comprehensive, unified view of segmentation. Unlike traditional approaches that focus solely on one aspect, panoptic segmentation categorizes visual entities into two classes: “stuff”and “thing”. The former encompasses amorphous regions like grass, sky, and road, suited for semantic segmentation, while the latter includes countable objects like people, animals, and tools, ideal for instance segmentation. By utilizing a predefined set of semantic classes divided into these two subsets, each pixel in an image is mapped to a pair consisting of a semantic label and an instance ID, enabling a uniform evaluation metric across all classes. This aligns perfectly with the objective of the thesis, which aims to incorporate all BE features depicted in images and transform them into suitable output formats. Hence, the semantic computer vision model in this thesis will leverage panoptic segmentation, where pixel-level categories correspond to “stuff”and instance-level categories to “thing”categories.

2.6. Image use in discrete choice modelling

Whether using images in stated choice modelling improves model performance remains inconclusive (Arellana et al. 2020). For instance, the studies of Iglesias et al. (2013) and Rossetti et al. (2018), which investigate safety perception regarding neighbourhoods and cycling infrastructures, found that image use can help respondents understand surveys and obtain better results of model parameters, respectively. However, studies that focus on transport mode choice (Arentze et al., 2003) and crowding discomfort in public transport (Tirachini et al., 2017) did not find evidence proving that image use can assist in more accurate data collection. Overall, in these studies, image use only assists in explaining the textual descriptions in the surveys auxiliarily rather than playing a leading role. One likely reason could be that researchers with subjective opinions select and extract attributes from images (Poudel & Singleton, 2022). van Cranenburgh and Garrido-Valenzuela (2023) demonstrates that the proposed CV-DCM (computer vision-enriched discrete choice model) extracting information embedded in images outperforms the traditional discrete choice model without images. This study suggests a practical approach for capturing information embedded in images and integrating it with discrete choice modelling by computer vision.

2.7. Other attributes of images affecting choices

Low-level features are introduced to characterize image attributes, encompassing HLS colour statistics (mean and standard deviation for hue, saturation, and lightness channels) and image edges (expressed as the percentage of pixels identified as edges). These variables are crucial for mitigating lighting and saturation variations, preventing biases in evaluations such as comparing images captured under different weather conditions. Edge statistics serve as a proxy measure for scene complexity, where unoccupied areas exhibit fewer edges than densely populated ones. Additionally, textured or detailed surfaces yield higher edge counts (Ito & Biljecki, 2021; Ramírez et al., 2021; Rossetti et al., 2019).
The season depicted in the image also influences individuals' choices. For instance, images captured during autumn and winter, when leaves fall, may not enhance the perceived utility of a residence. Conversely, images taken during spring or summer, with lush trees, may make the residence more appealing (Zhao et al., 2023). Moreover, research by van Cranenburgh and Garrido-Valenzuela (2023) demonstrates that incorporating the month into choice modelling improves prediction accuracy compared to solely including numeric attributes like housing costs and commuting time.

2.8. Takeaways

Previous studies on micro-scale BE features and RLC are few due to the difficulty of quantifying micro-scale BE features. Recent research on people's well-being and BE provide an emerging method on how to quantify micro-scale BE features: the emerging combination of semantic computer vision models and SVI.

Studies in Table 2.3 utilizing the Place Pulse 2.0 dataset offer valuable insights into the impacts of different built environment (BE) features on various perceptions. However, they fail to provide comprehensive insights on how to design a residential area considering all these perceptions since each perception of a neighbourhood plays a different weighted role in residential location choice (RLC). van Cranenburgh and Garrido-Valenzuela (2023) conducted a stated choice experiment wherein respondents were tasked with selecting a residence to live in between two alternatives, considering factors including images of residences, housing costs and commute travel time. This stated choice dataset presents additional opportunities to delve into the impacts of micro-scale BE features on RLC. By incorporating factors beyond visual perceptions, such as housing costs and commute time, this research avenue offers a more holistic understanding of the complex decision-making process of residential location choices.

Additionally, in these studies, BE features lack suitable categorization of semantic computer vision tasks, with some better suited for pixel-level quantification and others for instance-level quantification. Applying panoptic segmentation can incorporate all micro-scale BE features in suitable quantified units. This semantic computer vision task perfectly matches the research objective: a comprehensive understanding of the quantified impacts of all these features on people's preferences for residential neighbourhoods.

Besides extracting the micro-scale BE features from images, this thesis will also include season and low-level features of images to make the model capture more information people perceive.

3 Methodology

This chapter explains the methodology that is performed in this study in order to achieve the research objective. The following subsections include the introduction to the datasets to be used, the methodology framework, and a detailed description regarding two essential models in the methodology.

3.1. Datasets

As mentioned in section 1.1, this thesis will use the same datasets (choice task and SVI datasets) provided by van Cranenburgh and Garrido-Valenzuela (2023). The SVI dataset contains 7594 street-view images of residential streets in the Netherlands, retrieved from Google. The choice task dataset stems from a stated preference experiment conducted in September 2022. In this experiment, participants select one of two residential options which provide information on their numerical attributes (commuting time and housing costs) and images from the SVI dataset. The target population for the survey was the Dutch population of 18 years and older, with ten or more minutes of commute travel time. In total, 800 people participated in this experiment, each completing 15 choice tasks.

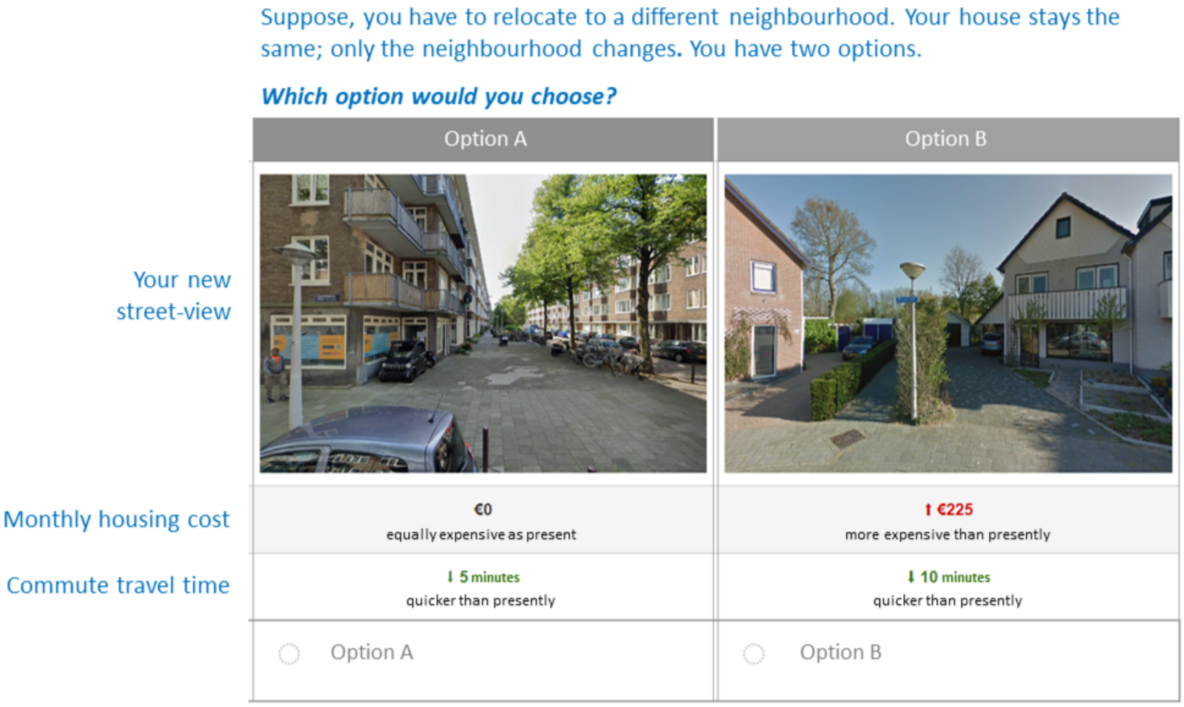
3.1.1. Street-view images

Each alternative in the dataset is accompanied by a street-level image, randomly sampled from a database they meticulously constructed beforehand. They initiated this process by randomly selecting 50 municipalities out of approximately 350 in the Netherlands to ensure a representative sample. Within these municipalities, they established a grid of points with 150-meter intervals in residential areas. Using Google's API, they retrieved street-view image IDs for each grid point, limited to images taken in 2020 or later. From each 360-degree panorama, they generated two image URLs, providing 90-degree angles to ensure one represents ”window views” (e.g., opposed to views parallel to the driving direction of the Google car). The algorithm removed images of poor quality (e.g., black and blurred ones), resulting in a database of over 60,000 street-view images. Notably, they recorded the month of capture for each image, recognizing the seasonal variations in the Netherlands. Given the distinct seasons and potential differences in environmental conditions, it is essential to consider how these factors might influence respondents' perceptions and preferences in the models.

3.1.2. Stated choice experiment

Respondents are prompted to envision themselves moving to a new neighbourhood and selecting their preferred residence from two options, each accompanied by street-view images and numerical attributes (housing cost and commuting travel time). Prior to beginning the choice experiment, respondents are informed of the following: 1) The new house mirrors their current one in terms of size, type, year built, furnishings, etc., with only the neighbourhood changing. 2) Monthly housing costs (including rent, mortgage, taxes, insurance, etc.) may fluctuate. 3) The new neighbourhood is relatively close to their current one, but commute time may vary. 4) All other aspects of their situation remain constant, including distances to amenities, schools, and healthcare providers. 5) The images presented in the choice tasks depict the street-level window view.

Reasons for adding numeric attributes:
The alternatives feature two key numeric attributes: monthly housing costs (hhc) and commute travel time (tti). They selected these attributes for their significance in residential location choice, broad applicability and usefulness in interpreting empirical results. By combining cost and time attributes, they can calculate the Value-of-Travel-Time (VTT), a widely studied metric in transport, aiding model validation. They deliberately limited attributes to demonstrate the proposed CV-DCMs' effectiveness in capturing visual preferences rather than developing an exhaustive predictive model for residential location choices. The experimental design, illustrated in Figure 3.1, employs a pivoted approach to present realistic choice scenarios, accounting for the variations in respondents' current situations, particularly housing costs. For housing cost, they utilized seven pivoted levels, while the number of levels and ranges for travel time depended on respondents' current travel times, as detailed in Table 3.1. These attribute ranges were determined through a small pilot study conducted prior to the main survey.



Suppose, you have to relocate to a different neighbourhood. Your house stays the same; only the neighbourhood changes. You have two options.

Which option would you choose?

	Option A	Option B
Your new street-view		
Monthly housing cost	€0 equally expensive as present	↑ €225 more expensive than presently
Commute travel time	↓ 5 minutes quicker than presently	↓ 10 minutes quicker than presently
	○ Option A	○ Option B

Figure 3.1: Screenshot of a choice task from van Cranenburgh and Garrido-Valenzuela (2023)

Table 3.1: Attribute levels stated choice experiment: from van Cranenburgh and Garrido-Valenzuela (2023)

Current commute travel time of the respondent (TT_n)	Attribute levels	
	Housing cost (hhc) [€]	Commute travel time (tti) [min]
$TT_n < 10min$	N/A	
$10min < TT_n < 20min$	-225, -150, -75, 0, +75, +150, +225	-5, 0, +5, +10, +15
$20min < TT_n < 30min$		-10, -5, 0, +5, +10, +15
$10min < TT_n$		-15, -10, -5, 0, +5, +10, +15

Design of each each choice tasks:
Figure 3.1 indicates their choice of a pivoted experimental design to provide respondents with realistic choice scenarios. Instead of absolute-level designs, they opted for this pivoted-level design to address the considerable variations in respondents' current situations. Particularly, they utilized seven pivoted levels for the housing cost attribute and adjusted the number of levels and ranges for the travel time attribute based on each respondent's current travel time, as detailed in Table 3.1. A preliminary pilot study conducted before the main survey determined the ranges for both attributes.

They employed a random experimental design due to the absence of ordinal or categorical levels in the images, making it unfeasible to adopt an orthogonal or efficient experimental design strategy, especially considering the images. Thus, they followed a two-step approach in constructing the choice tasks. First, they randomly selected image pairs, ensuring none were from respondents' municipalities identified by their provided postcodes. This aimed to reduce biases from respondents' familiarity with local images; though complete unfamiliarity was not guaranteed, it decreased the likelihood.

Next, the researchers introduced housing cost (hhc) and travel time (tti) levels into the choice tasks. This was done by randomly selecting a choice task from one of three pre-generated tables. Each table was created by first establishing a full-factorial design based on the attribute levels outlined in Table 3.1. Then, they eliminated choice tasks that didn't involve a trade-off between housing costs and travel time, guided by strong prior beliefs about the expected sign of the preference parameters for these attributes. This step was crucial in ensuring that the choice tasks were aligned with the research objectives. Lastly, they excluded choice tasks where attribute levels were identical, a measure that was taken to ensure that each task inherently entailed a trade-off between housing cost and travel time, a key aspect of the research study.

3.2. Methodology workflow

Figure 3.2 illustrates the methodology workflow of this study, comprising two main components: the semantic computer vision model and the discrete choice model (DCM). The semantic computer vision model will leverage the Panoptic-Segment-Anything model ((PSAM) proposed by Tobias Cornille [1] to quantify micro-scale BE features. Concurrently, the multinomial logit model will be adopted for choice modelling, consistent with the methodology employed in van Cranenburgh and Garrido-Valenzuela (2023).

[1] https://github.com/segments-ai/panoptic-segment-anything

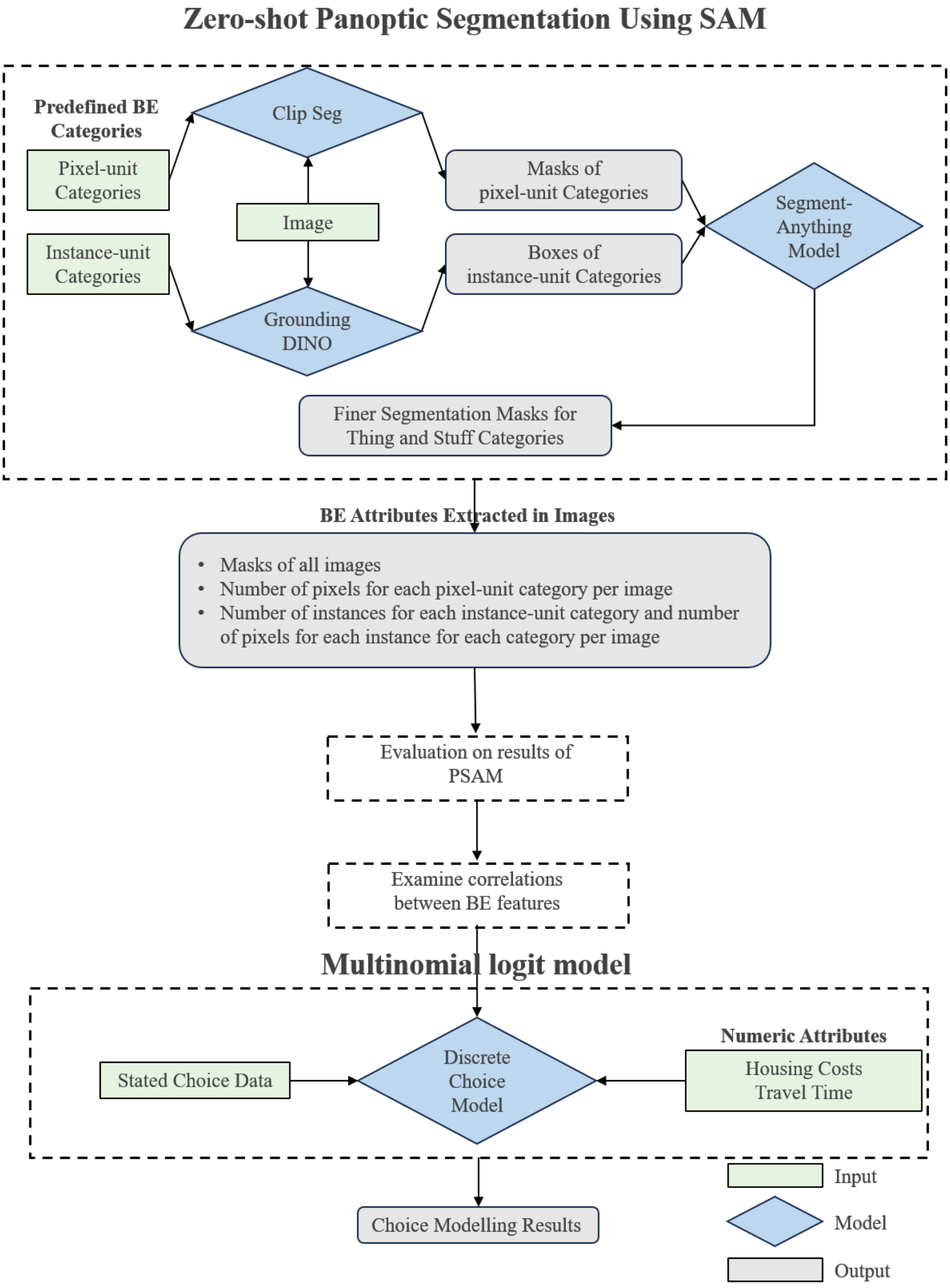

Zero-shot Panoptic Segmentation Using SAM
Predefined BE Categories
Pixel-unit Categories
Instance-unit Categories
Clip Seg
Image
Grounding DINO
Masks of pixel-unit Categories
Boxes of instance-unit Categories
Segment-Anything Model
Finer Segmentation Masks for Thing and Stuff Categories
BE Attributes Extracted in Images
Masks of all images
Number of pixels for each pixel-unit category per image
Number of instances for each instance-unit category and number of pixels for each instance for each category per image
Evaluation on results of PSAM
Examine correlations between BE features
Multinomial logit model
Stated Choice Data
Discrete Choice Model
Numeric Attributes
Housing Costs
Travel Time
Choice Modelling Results
Input
Model
Output

Figure 3.2: Methodology workflow

The inputs for the semantic computer vision model encompass street-view images depicting the surrounding built environment (BE) of residences and semantic texts representing micro-scale BE features, further classified as pixel-unit and instance-unit categories. To predefine micro-scale BE features, pertinent literature such as Ramírez et al. (2021) and Rossetti et al. (2019) can provide insights into commonly considered visual elements. Additionally, manual inspection of image data beforehand is crucial to record frequently appearing elements. The PSAM will output masks, the number of pixels for each pixel-unit category, and the number of instances for each instance-unit category, including the number of pixels for each instance within each image.

Following PSAM results, assessing their accuracy for subsequent steps is imperative. Evaluators will select a subset of images, locate their corresponding generated masks and numerical results, and estimate the proportion of correctly identified pixels for each image. Mask evaluation ensures the reliability of image processing, with only sufficiently accurate results proceeding to the following steps. Moreover, assessing the generated masks can identify missed or frequently misidentified micro-scale BE features, prompting updates to predefined categories and potentially rerunning the workflow to enhance accuracy. Simultaneously, reclassifying challenging micro-scale BE features into alternative categories could further refine results, necessitating multiple workflow experiments in such cases.

Consistently observed features in street-view images, such as interconnected roadways and sidewalks or the presence of trees and sky, may exhibit correlations in their quantified values. These potential correlations can complicate individual attributions and challenge statistical inference. High correlations can significantly affect coefficient estimation accuracy. Therefore, examining correlations between quantified values of micro-scale BE features is crucial.

After evaluating masks and feature correlations, the quantified micro-scale BE features will be integrated into the multinomial logit model. Image and stated choice data are sourced from van Cranenburgh and Garrido-Valenzuela (2023), where participants selected one of two residential options in a stated preference experiment based on numerical attributes (commute travel time and housing costs) and corresponding images. Consequently, commute travel time and housing costs will serve as variables in the choice modelling, yielding an estimating model with BE feature variables and numeric attributes, including model performance metrics such as log-likelihood ratio.

3.3. Panoptic-Segment-Anything Model

The Panoptic-Segment-Anything model (PSAM) is a zero-shot panoptic segmentation model using the Segment Anything Model (SAM). Panoptic segmentation offers a unified framework that combines instance and semantic segmentation to provide a comprehensive understanding of visual scenes, distinguishing between “thing”(i.e., instance-unit) and “stuff”(i.e., pixel-unit) categories.

The Segment Anything Model (SAM), developed by Meta AI as a novel model, offers a powerful and versatile solution for image object segmentation. It applies to various semantic computer vision tasks, including object detection, semantic segmentation and instance segmentation (Kirillov et al., 2023). However, SAM cannot immediately achieve panoptic segmentation, which is attributed to the fact that the released version of SAM is not text-aware. In other words, the SAM cannot take texts as prompts to identify and classify regions into corresponding semantic categories.

To solve these challenges, Tobias Cornille uses the following additional models: Grounding DINO (Liu et al., 2023), a zero-shot object detector, and CLIPSeg (Lüddecke & Ecker, 2022), a zero-shot (binary) segmentation model. The pipeline is demonstrated in the upper part of figure 3.2. Firstly, they use Grounding DINO to detect the instance-unit categories. Secondly, instance segmentation masks for the detected boxes will be obtained using SAM. Thirdly, CLIPSeg will obtain rough segmentation masks of the pixel-unit categories. Fourth, sample points in these rough segmentation masks and feed these to SAM to get fine segmentation masks. Last, combine the background pixel-unit masks with the foreground instance-unit masks to obtain a panoptic segmentation label.

3.4. The evaluation metric on results of PSAM

While studies that apply semantic computer vision models for downstream tasks often overlook the evaluation of the generated masks (Ramírez et al., 2021; Rossetti et al., 2019), this thesis proposes a novel method for assessing the accuracy of semantic computer vision models in labelling and segmenting images. This method addresses the common belief among scholars that segmentation performances are adequately high but also acknowledges the lingering question of whether a trained semantic computer vision model can accurately label pixels and objects in a new dataset. The following evaluation metric provides a convincing approach with substantial evidence.

The standard evaluation metric for panoptic segmentation is the panoptic quality (PQ), which involves two steps: 1) segment matching and 2) PQ computation given the matches (More details see Kirillov et al. (2019)). However, due to the lack of ground truth masks of all pixels in the SVI dataset, the evaluation metric PQ proposed for training the model is complicated for humans to follow. Also, assessing the generated masks of a zero-shot model does not necessarily require highly precise evaluation metrics since the aim is to demonstrate whether the model performance is sufficiently accurate for downstream tasks (i.e., choice modelling). Thus, the PQ will not be adopted as the evaluation metric in the following contents.

To assess whether the PSAM correctly labels and segments regions in images in a relatively easy way, the evaluation metric for each mask is divided into two steps. 1) Record the incorrectly labelled instances for all instance-unit and incorrectly labelled proportions of pixels for all pixel-unit categories, respectively. 2) Take a sum of all incorrectly labelled regions in the unit of pixel proportion and assess the boundaries between segmented regions based on own subjective opinions. It is worth noting that each image is to be evaluated as an Excel file recording the proportion of pixels for each identified pixel-unit category and the number of instances along with its proportion for each identified instance-unit category.

The first step focuses on the labelled number of instances for instance-unit categories and the proportions of labelled pixels for pixel-unit categories as they are two forms of data to be fed to the following discrete choice models, which are essential to be assessed. Since the correctly identified regions still account for more than the incorrectly identified ones for most generated masks, the number of incorrectly identified instances for instance-unit categories (i.e., II_thing_instance) and the proportion of incorrectly identified pixels for all pixel-unit categories (i.e., II_stuff_pixel) will be roughly estimated by humans based on the output excel file per image. Note that the estimations on the pixel proportions are all calculated by the sizes of the regions based on the coordinate axis of images.

The second step aims to examine the general accuracy of the generated mask for each image. It considers two aspects: the overall accurate assignment of semantic categories to pixels and the precision of boundaries between segmented regions. The former evaluates the model's ability to label regions in images with semantic categories correctly. It is similar to Pixel Accuracy (PA), a standard evaluation metric for semantic segmentation that calculates the percentage of correctly labelled pixels over the total number of pixels in the image (Kirillov et al., 2019). The evaluator estimates the proportion of pixels for all incorrectly identified instances (i.e., II_thing_pixel) and sums it with recorded II_stuff_pixel to have the total proportion of incorrectly identified pixels (i.e., II_P.P), as in equation 3.1.

$$II\_P.P = II\_thing\_pixels + II\_stuff\_pixels \tag{3.1}$$

$$II\_thing\_instance = abs(G\_thing - I\_thing) \tag{3.2}$$

Regarding the assessment of boundaries, each image will receive a subjective score indicating the precision of segmentation. Also, it remedies the discontinuous regions of which proportions are hard to estimate by humans. Assessing the overall quality of boundaries between segmented regions may help address the issue. The predefined scores include 0, 1, 2, and 3, corresponding to unclear, relatively unclear, relatively clear, and clear boundaries.

3.5. Discrete Choice Model

The bottom of Figure 3.2 outlines two primary types of variables essential for inclusion in choice modelling: BE attributes extracted from images and numeric attributes (housing costs and commuting time). Similar to the computer-vision discrete choice model proposed by van Cranenburgh and Garrido-Valenzuela (2023), there are several assumptions in the multinomial logit model.

1) Decision-makers are assumed to make decisions based on Random Utility Maximising (RUM) principles, see Equation 3.3, with U_{jn} denoting the total utility experienced by decision-maker n considering alternative j, V_{jn} is the utility experienced by decision-maker n derived from attributes observable by the analyst. Also, an additive error term ε_{jn} is added to each alternative because the analyst does not observe everything that matters to the decision-makers's utility.

$$U_{jn} = V_{jn} + \varepsilon_{jn} \tag{3.3}$$

2) During respondents completing each choice task, the information provided to them are images ((S_{jn}) of two residences and their corresponding numeric attributes (X_{jn}, represent monthly housing costs and commuting travel time). Equation 3.4 shows that v is a preference function which maps the attributes and image onto the utility.

$$U_{jn}(X_{jn}, S_{jn}) = v(X_{jn}, S_{jn}) + \varepsilon_{jn} \tag{3.4}$$

3) The utility is derived from the numeric attributes, and the image is assumed to be separable and additive in utility space. Equation 3.5 shows that function z maps the numeric attributes onto the utility and function g maps the information from the image onto the utility.

$$U_{jn}(X_{jn}, S_{jn}) = z(X_{jn}) + g(S_{jn}) + \varepsilon_{jn} \tag{3.5}$$

4) Two supplementary variables will be considered: season and low-level features (LLF) of images since they significantly impact people's perceptions of images, as discussed in the literature review. Without differentiating LLF and season variations from images, the gained impacts of micro-scale BE features by choice modelling are possibly biased. It is worth including LLF and seasons into the discrete choice model to estimate more accurate coefficients of micro-scale BE features. Also, LLF encompassing HLS colour statistics (mean and standard deviation for hue, saturation, and lightness channels) and image edges (expressed as the percentage of pixels identified as edges) are easy to measure by libraries like OpenCV and NumPy in Python. Besides, the SVI dataset by van Cranenburgh and Garrido-Valenzuela (2023) records the month of all images, which can be further categorized into seasons.

Therefore, assuming information people perceive in images does not only include quantification of different micro-scale BE features but also LLF (e.g., lightness) and season variations (e.g., whether leaves of trees are present or not). The three types of information are linear additive, influencing people's choices. Model 1 to Model 4 are four models representing four utility functions, corresponding to the two required and two optional types of variables as outlined below. Model 0 is the benchmark model to compare with the proposed Model 1 to Model 4, which only incorporates housing costs and travel time.

Model 0: Numeric attributes:

$$U_{jn} = \sum_{m} \beta_m X_{jmn} + \varepsilon_{jn} \tag{3.6}$$

Model 1: Numeric attributes + micro-scale BE features:

$$U_{jn} = \sum_{m} \beta_m X_{jmn} + \sum_{f} \beta_f S_{jfn} + \varepsilon_{jn} \tag{3.7}$$

Model 2: Numeric attributes + micro-scale BE features + season:

$$U_{jn} = \sum_{m} \beta_m X_{jmn} + \sum_{f} \beta_f S_{jfn} + \sum_{h} \beta_h S_{jhn} + \varepsilon_{jn} \tag{3.8}$$

Model 3: Numeric attributes + micro-scale BE features + Low-level features (LLF):

$$U_{jn} = \sum_{m} \beta_m X_{jmn} + \sum_{f} \beta_f S_{jfn} + \sum_{l} \beta_l S_{jln} + \varepsilon_{jn} \tag{3.9}$$

Model 4: Numeric attributes + micro-scale BE features + Season + Low-level features (LLF):

$$U_{jn} = \sum_{m} \beta_m X_{jmn} + \sum_{f} \beta_f S_{jfn} + \sum_{h} \beta_h S_{jhn} + \sum_{l} \beta_l S_{jln} + \varepsilon_{jn} \tag{3.10}$$

Where U_{jn} denotes the total utility experienced by decision-maker n considering alternative j; V_{jn} is the utility experienced by decision-maker n derived from attributes observable by the analyst; m refers the numeric attributes (i.e. housing costs, commuting travel time); f represents different micro-scale BE features; h refers the four seasons; l represents the seven low-level features (LLF)

5) Following the typical approach in choice modelling, it is assumed that ε_{jn} is independently and identically distributed according to the Extreme Value Type I distribution, with a variance of $\pi^2/6$This assumption leads to the widely recognized and easily computed closed-form logit formula for the choice probabilities P_{in}, as detailed in Equation 3.11.

$$P_{in} = \frac{e^{V_{in}}}{\sum\limits_{j \in C_n} e^{V_{jn}}} \tag{3.11}$$

Where C_n represents the set of alternatives presented to decision maker n.

3.5.1. Normalization

After applying the semantic computer vision model, each image contains quantified values representing micro-scale BE features alongside extracted low-level features (LLF). Notably, pixel-unit features range from 0 to 1, while instance-unit features consist of non-negative integers. The ranges of different LLF also vary. To ensure an accurate comparison of variable estimation parameters in choice modelling, it is essential to normalize these diverse variable types. This normalization process will utilize Min-Max Normalization.

Pixel-unit categories across all images will undergo normalization to maintain disparities between different pixel-unit categories in one image and disparities of the same pixel-unit category between differing images. Instance-unit categories will be applied the same way as pixel-unit categories to keep the disparities. For LLF, normalization will be straightforward, with each feature normalized across all images.

3.5.2. Strategy searching for valid model specification

In pursuit of a model specification ensuring the significant influence of all variables within the utility function on choice behaviour (i.e., at least 90% confidence), a build-down strategy is adopted for each type of utility function. For instance, when refining the model specification for Model 1, all numeric attributes and categories are initially included in the utility function, followed by the systematic removal of variables surpassing the predetermined p-value threshold from highest to lowest. Throughout this iterative process, model performance indicators such as the log-likelihood ratio consistently fluctuate within a narrow range. Once a model is attained where all variables satisfy the t-test/p-value criterion, some previously removed variables are reintroduced to assess if they can meet the t-test threshold.

In most cases, reintroduced variables fall short of satisfying the t-test criterion. While this build-down strategy may not explore all possible variable compositions, it enables the ascertainment of a valid model specification guaranteeing the significant influence of all included variables on choice behaviour. The motivation behind incorporating additional variables is to address the research question effectively. By including more micro-scale BE features in the model specification, we can explore the impacts of a broader range of features on residential location choice.

4
Results

This section analyzes the results of the PSAM and DCM after an explanation on the selected BE features as inputs to be identified by the PSAM. The subsequent sub-sections encompass the PSAM's outcomes, comprising how to apply the evaluation metric, the evaluation results, and descriptive statistics detailing the quantified BE features across all images. An examination of correlations between the quantified BE features follows this. Subsequently, the results of the discrete choice models and an analysis of the estimated coefficients are presented.

4.1. Selected micro-scale BE features

Identifying relevant environmental features poses a significant challenge in studies examining people's perceptions or behaviours (Zhanjun et al., 2022). Tables 2.2 and 2.3 showcase micro-scale BE features that previous scholars have selected for investigation. Given their contextual relevance to this thesis, these features serve as valuable references. However, it is crucial to acknowledge that certain features listed in these tables may not be prevalent in the Netherlands' SVI dataset. Therefore, manual inspection of some SVI images is necessary to identify frequent features.

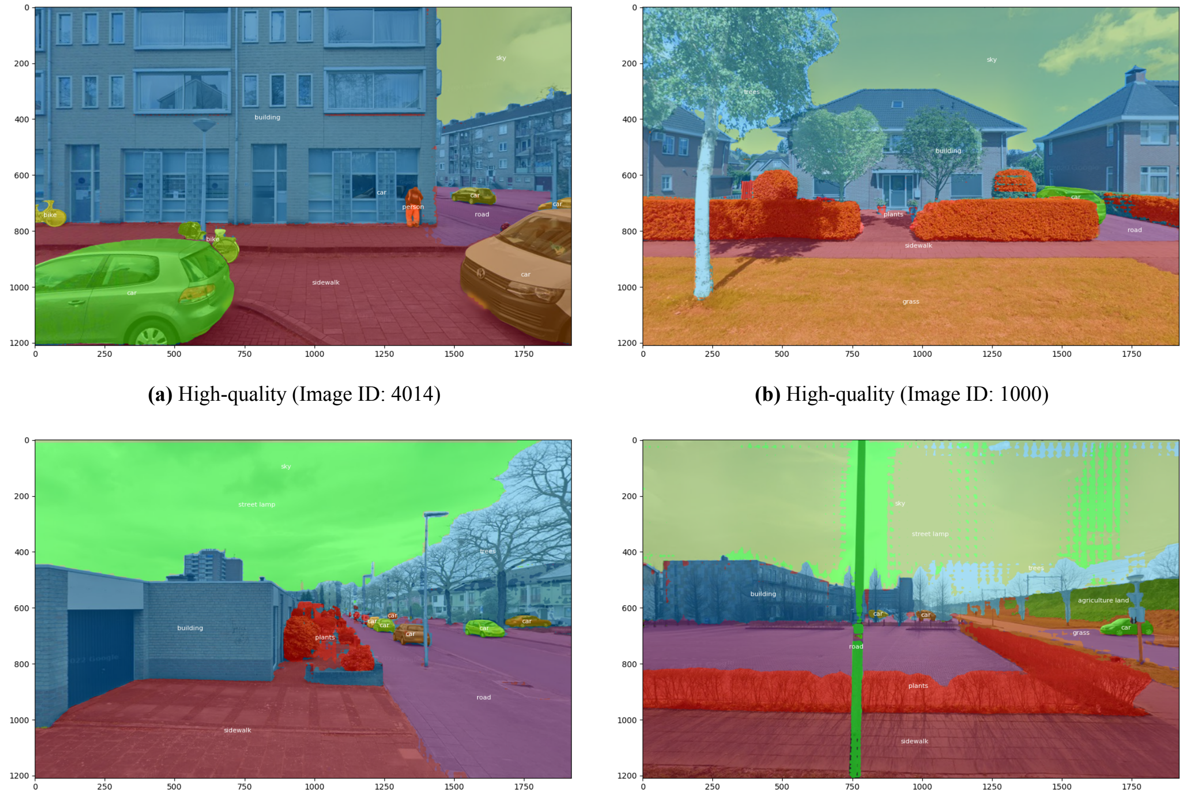
Following reviewing around 200 images of the SVI data, BE features that have a frequency more than 1 are selected as inputs for PSAM, forming a category list and presenting in Table 4.1. Most micro-scale BE features are anticipated to impact RLC in specific ways. For instance, vegetation and water may positively influence RLC by enhancing the aesthetic appeal of surroundings. At the same time, amenities like bus stops and dustbins may deposit positive affects on RLC because they facilitate convenient access to transportation and essential services. Conversely, factors such as motorcycles may negatively impact RLC due to associated noise disturbances.

Table 4.1: Predefined micro-scale BE features in category lists

Category list	Micro-scale BE features
Pixel-unit Categories	building, grass, road, sky, trees, sidewalk, plants, street sign, traffic light, fence, street lamp, water, fire,hydrant, distribution box, agriculture land
Instance-unit Categories	car, person, bench, dustbin, boat, bike, motorcycle, bus stop

4.2. Evaluation on results of PSAM

A random selection of 400 images from 7653 images is used for the segmentation quality/performance evaluation. Figure 4.1 showcases four examples of masks generated by the PSAM. The top two illustrate high-quality masks, where most BE features are accurately assigned predefined semantic categories, and distinct boundaries between features are evident. Conversely, the bottom two display low-quality masks due to incorrect category assignments and unclear boundaries. The comparison between high- and low-quality masks show that it is essential to randomly evaluate part of the generated masks to gauge the overall quality and verify the accuracy of the PSAM's results for choice modelling purposes.



(a) High-quality (Image ID: 4014)

(b) High-quality (Image ID: 1000)

(c) Low-quality (Image ID: 5679)

(d) Low-quality (Image ID: 4980)

Figure 4.1: Comparison between high- and low-quality segmentation masks

4.2.1. Applying the evaluation metric

Table 4.2 shows the evaluation results for the above four images. Take Image ID 4014 (i.e., Figure 4.1a) as an example; the evaluator records the number of ground-truth instances (i.e., G_thing) as 7. The number of instances (i.e., I_thing) identified by the model is 8 in the Excel file of the mask for Image ID 4014. Thus, the incorrectly identified instance for the image is 1, which is the car reflected in the mirror, as calculated in equation 3.2. II_thing_pixels is 0.001 based on the proportion of pixels for the car in the Excel file of the mask. The evaluator records the number of ground-truth pixel-unit categories (i.e., G_thing) and compares it with the number of identified pixel-unit categories (i.e., I_thing). As there are five ground-truth categories but four identified categories, the evaluator finds out the missing one is a street lamp and estimates the size of the corresponding region. Hence, II_stuff_pixel is calculated as 0.001 by dividing the region's size by the image size. II_P.P is the summation of II_thing_pixels

(0.001) and II_stuff_pixel (0.001): 0.002. The boundaries between segmented regions remain clear, so this image receives a high score of 3.

Table 4.2: Evaluation table of the example figure

Image ID	4014	1000	5679	4980
G_thing	7	1	6	3
G_stuff	5	7	7	9
I_thing	8	1	6	3
I_stuff	4	7	7	7
II_thing_instance	1	0	0	0
II_thing_pixel	0.001	0.000	0.000	0.000
II_stuff_categories	1	0	1	2
II_stuff_pixel	0.001	0.000	0.300	0.099
II_P.P	0.002	0.000	0.300	0.099
Boundary	3	3	2	2

G_thing: Number of ground-truth instances of instance-unit (thing) categories
G_stuff: Number of ground-truth pixel-unit (stuff) categories
I_thing: Number of instances of instance-unit (thing) categories that identified by the model
I_stuff: Number of pixel-unit (stuff) categories that are identified by the model
II_thing_instance: Number of thing instances that are incorrectly identified by the Model
II_thing_pixels: Proportion of pixels of thing instances that are incorrectly identified by the model
II_stuff_categories: Number of pixel-unit (stuff) categories that are incorrectly identified by the model
II_stuff_pixels: Proportion of pixels of pixel-unit (stuff) categories that are incorrectly identified by the model
II_P.P: Proportion of pixels that by incorrectly identified by the model (rough estimation)
Boundary: Subjective evaluation of the model's segmentation precision. Values from 0-3, the higher the value, the clearer boundaries.

4.2.2. Evaluation results

After analyzing 400 randomly selected images, Figures 4.2 and 4.3 illustrate the outcomes of the first step in the evaluation metric: the histogram showcasing incorrectly identified instances for instance-unit categories and the proportions of incorrectly identified pixels for pixel-unit categories, respectively.

Among the images, 72.8% exhibit zero incorrectly identified instances, while 15.0% (89.8% - 72.8%) have one misidentified instance. Approximately 10% of the images show more than two instances incorrectly identified. Concerning pixel-unit categories, Figure 4.3 reveals that 27.6% of the images display 0% incorrectly identified regions, and 76.9% have less than 10% of pixels incorrectly identified.

Both the two histograms exhibit similar trends. Both curves peak at around 75% when the x-axis value is small, indicating that for 75% of the images, the number of incorrectly identified instances and the proportion of misidentified pixels are 0 and less than 10%, respectively. The two histograms suggest that for most images, the predicted quantities for each category closely align with the ground-truth. The PSAM's results can be utilized for subsequent choice modelling processes.

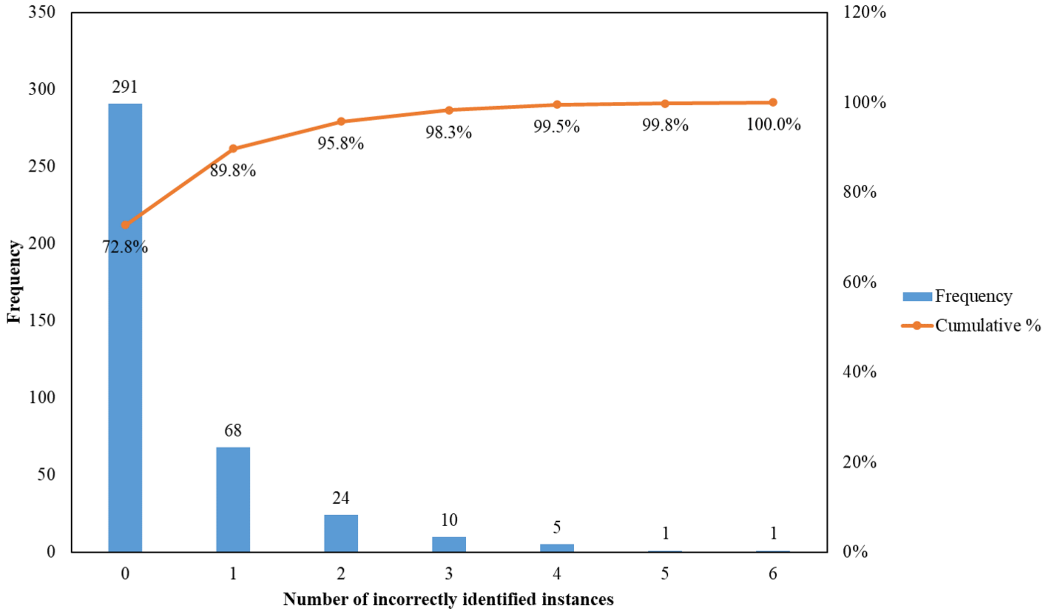


Figure 4.2: Histogram of II_thing_instance

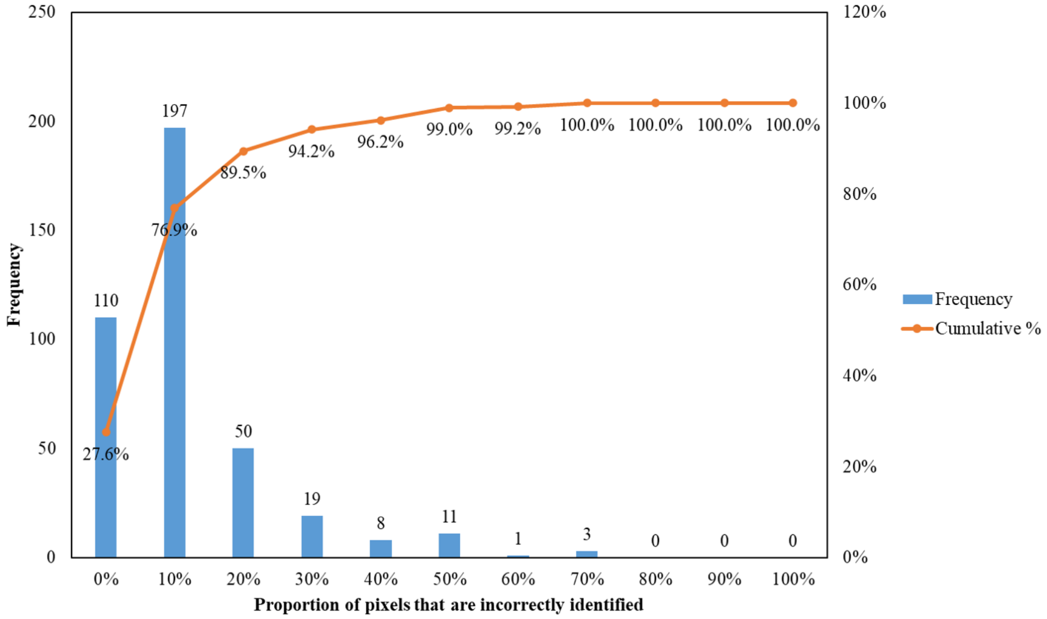


Figure 4.3: Histogram of II_stuff_pixel

Figure 4.4 and 4.5 showcase the outcomes of step 2 in the evaluation metric, illustrating histograms of II_P.P and boundaries. Among the images, 20% display 0% incorrectly identified pixels, while 74.8% exhibit less than 10% of pixels being inaccurately identified. Regarding boundaries, only 14.5% of the images are associated with values of 1 and 0, indicating somewhat ambiguous and disorderly boundaries, respectively, suggesting that the remaining images feature relatively clear boundaries.

Figure 4.4 demonstrates a generally accurate assignment of semantic categories to all pixels across all images. Meanwhile, Figure 4.5 showcases the high precision of boundaries between segmented regions. When the two histograms are examined together, they collectively portray satisfactory segmentation outcomes, affirming the overall effectiveness of the results of the PSAM.

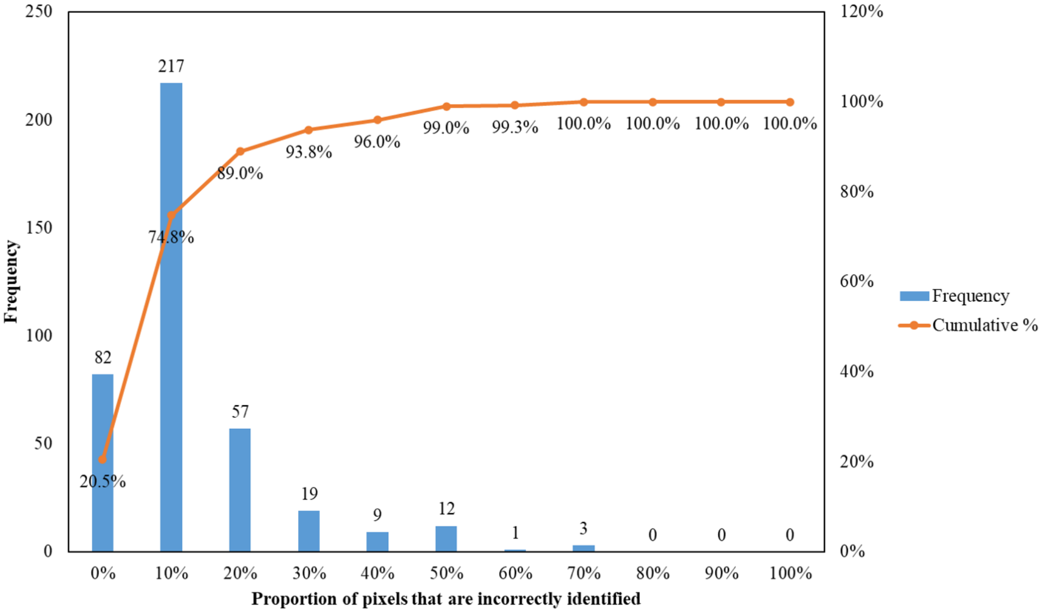


Figure 4.4: Histogram of II_P.P

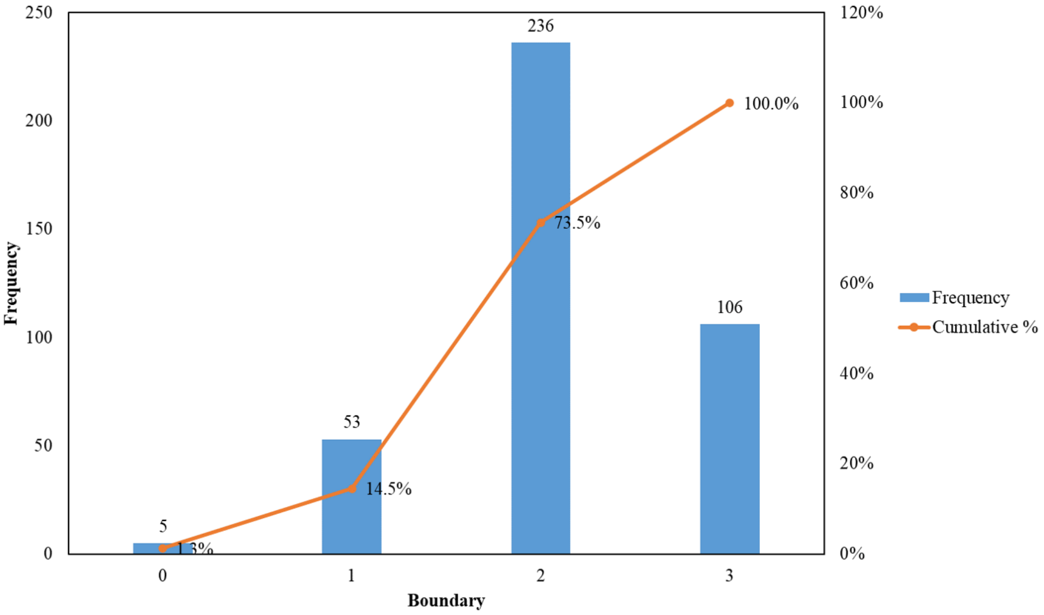
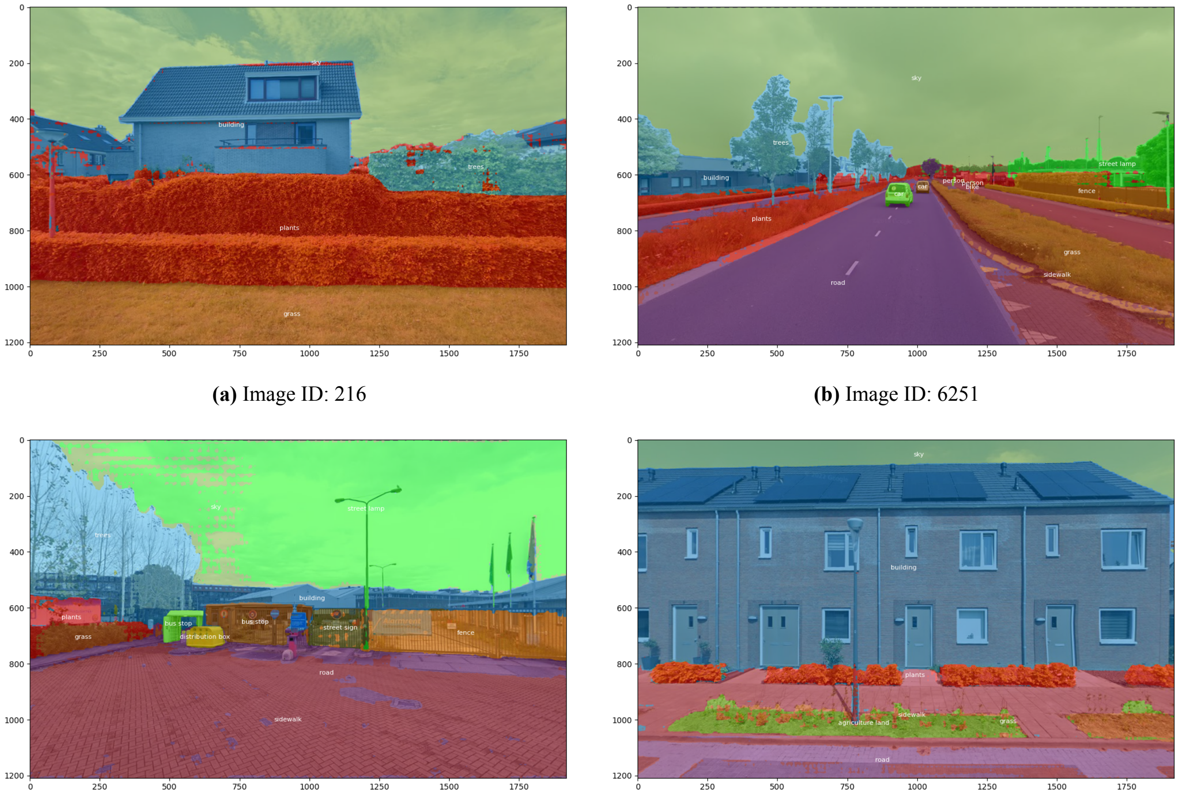


Figure 4.5: Histogram of Boundary

4.2.3. Specific categories

During the evaluation process, the evaluator also documents essential notes for each generated mask, such as ”part of grass identified as plants” or ”part of unknown regions identified as a bus stop.” After assessing all 400 masks, the evaluator identified common occurrences in the masks. The following discussion outlines key findings for specific categories.

In terms of identifying street lamps, the model faces a complex task. Smaller ones pose a significant challenge for accurate identification, whereas larger ones yield better performance, as depicted in Figure 4.6a. The model often struggles with misidentifying surrounding elements like vegetation or the sky as street lamps, as shown in Figure 4.6b and 4.6c. The complexity increases when buildings stand behind the lamp, making street lamp identification problematic, likely due to similar colours, as seen in Figure 4.6d.



(a) Image ID: 216

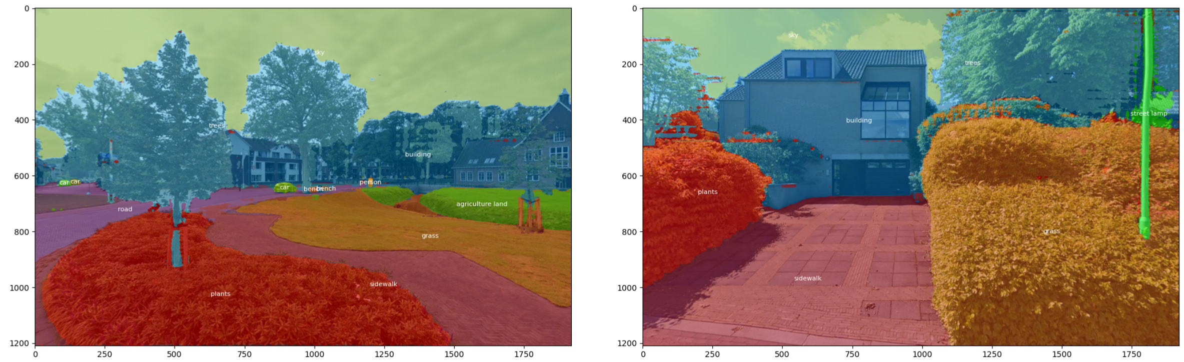
(b) Image ID: 6251

(c) Image ID: 6022

(d) Image ID: 6017

Figure 4.6: Examples of misidentified masks of street lamps

Vegetation is categorized into four types: trees, plants, grass, and agricultural land. However, the model struggles to distinguish them accurately. Grass and agricultural land are frequently mislabeled as each other, as are plants and grass, as demonstrated in Figures 4.10a and 4.10b, respectively. Conversely, trees are more consistently identified correctly than the other three vegetation types. Thus, the semantic accuracy of the PSAM for vegetation categories influences the credibility of their estimated parameters in discrete choice models. Interpreting the estimated coefficients for vegetation variables must consider their potential lack of differentiation in the PSAM.

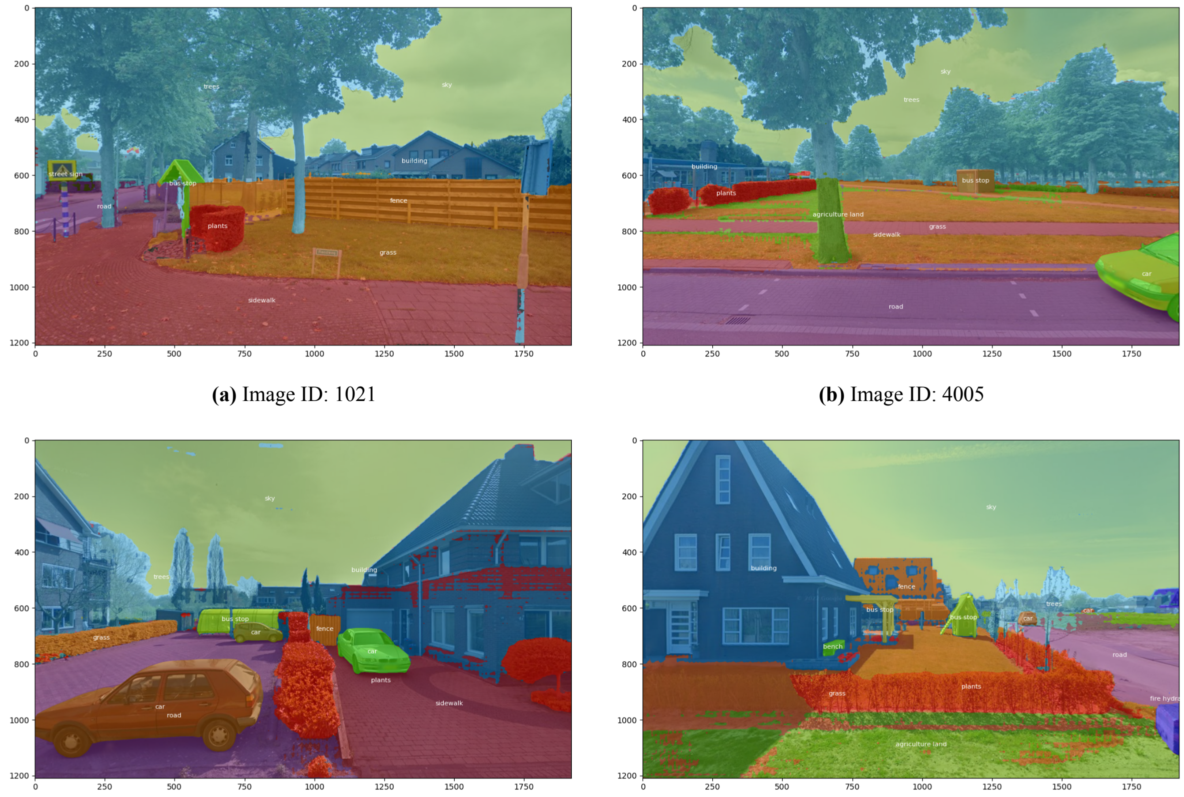


(a) Image ID: 4012, Grass identified as agriculture land

(b) Image ID: 4008, Plant identified as grass

Figure 4.7: Examples of misidentified masks of vegetation categories

In the case of bus stops, many labelled instances in the images do not correspond to actual bus stops. Regions marked as bus stops often encompass areas challenging to categorize, even for humans, as exemplified in Figure 4.8. These random, disordered, semantic-less regions often receive the bus stop label despite actual occurrences of bus stops numbering few among the 400 images. Consequently, the number of images featuring bus stops in the PSAM results could be relatively high.



(a) Image ID: 1021

(b) Image ID: 4005

(c) Image ID: 4006

(d) Image ID: 4017

Figure 4.8: Examples of misidentified masks of bus stops

Similarly, distribution boxes face a comparable issue. While the number of images featuring distribution boxes is relatively low, some semantic-less regions in many images are labelled as distribution boxes by the PSA model, as illustrated in Figure 4.9. The low labelling accuracy of these two categories (bus stops and distribution boxes) underscores the limitations of the zero-shot computer vision model in handling them.

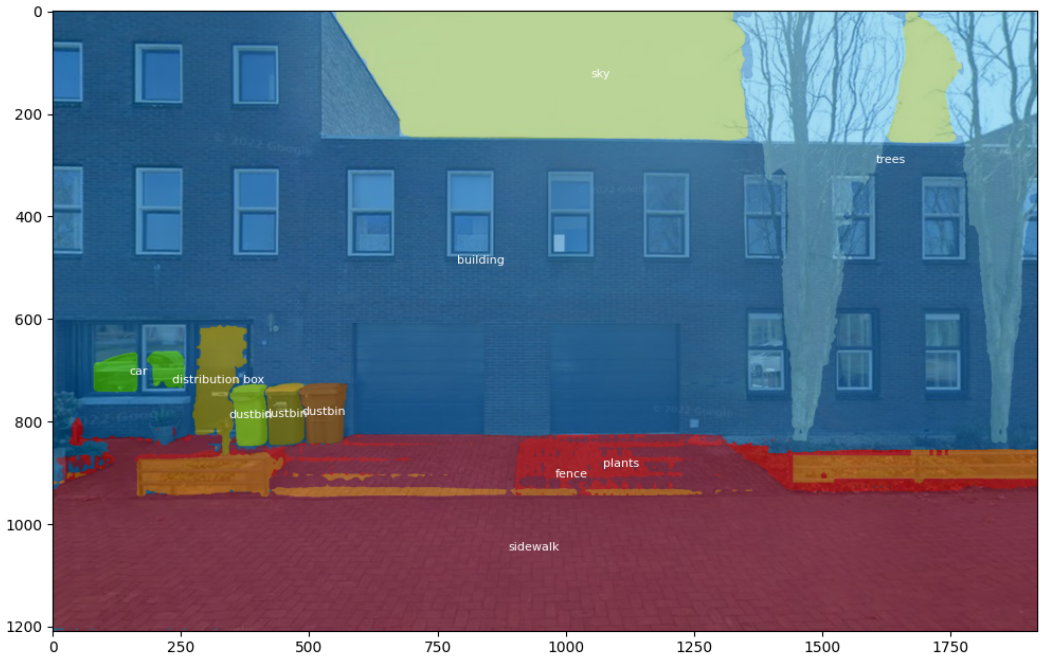
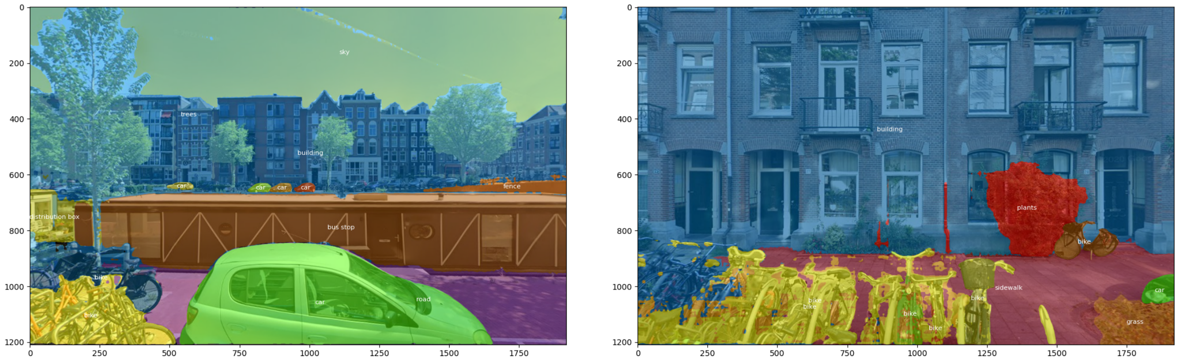


Figure 4.9: Examples of misidentified masks of distribution boxes

On the other hand, several categories, such as traffic lights, street signs, and fire hydrants, also seldom appear in the images. Unlike bus stops and distribution boxes, the PSAM results indicate that few images contain these categories. It may suggest that the PSAM handles these three categories better than bus stops and distribution boxes. Despite the evaluator noting misidentifications concerning these categories, the low frequency of occurrence mitigates potential issues arising from such misidentifications.



(a) Image ID: 6764

(b) Image ID: 7010

Figure 4.10: Examples of misidentified masks of bikes

Bikes also warrant attention due to their relatively frequent misidentification. The evaluator observed that bikes are often stacked instead of correctly placed in bike racks. It could contribute to fewer bikes identified than the actual count. Figure 4.10 provides an example where bikes are not fully and accurately identified when stacked together.

4.2.4. Descriptive statistics

Table 4.3 presents the descriptive statistics outlining the quantified BE features. It is important to note the distinction between pixel-unit and instance-unit categories: the former denotes percentages of pixels, while the latter represents counts of instances. These features are ordered based on their occurrence frequency. Elements such as sky, buildings, sidewalks, roads, and vegetation emerge as the most prevalent ones in images. In contrast, items like fire hydrants, traffic lights, water, and street signs are comparatively rare.

Table 4.3: Descriptive statistics

Categorization	BE features	Count	Mean	Std.	Min	25%	50%	75%	Max
		(1)	(2)	(2)					
Pixel-unit Categories	sky	7016	0.176	0.118	0.000	0.081	0.172	0.258	0.993
	building	6914	0.232	0.201	0.000	0.070	0.189	0.344	1.000
	trees	6646	0.157	0.148	0.000	0.031	0.116	0.250	0.957
	sidewalk	6303	0.095	0.084	0.000	0.023	0.077	0.150	0.542
	plants	5996	0.059	0.076	0.000	0.008	0.035	0.080	0.798
	grass	5990	0.087	0.099	0.000	0.010	0.053	0.133	0.791
	road	5122	0.074	0.087	0.000	0.000	0.042	0.124	0.553
	fence	3431	0.027	0.055	0.000	0.000	0.000	0.030	0.828
	agriculture land	2136	0.025	0.060	0.000	0.000	0.000	0.015	0.772
	street lamp	1195	0.009	0.040	0.000	0.000	0.000	0.000	0.692
	distribution box	1070	0.006	0.032	0.000	0.000	0.000	0.000	0.842
	water	562	0.006	0.032	0.000	0.000	0.000	0.000	0.430
	street sign	553	0.001	0.015	0.000	0.000	0.000	0.000	0.610
	traffic light	127	0.001	0.009	0.000	0.000	0.000	0.000	0.458
	fire hydrant	107	0.000	0.003	0.000	0.000	0.000	0.000	0.105
Instance-unit Categories	car	4560	1.823	2.518	0	0	1	3	21
	bus stop	1254	0.187	0.446	0	0	0	0	4
	bike	1066	0.315	1.050	0	0	0	0	14
	bench	879	0.164	0.536	0	0	0	0	10
	person	824	0.211	0.832	0	0	0	0	20
	dustbin	691	0.145	0.580	0	0	0	0	9
	motorcycle	260	0.047	0.300	0	0	0	0	8
	boat	111	0.040	0.437	0	0	0	0	11

An intriguing observation is the presence of images where a single category dominates the scene. For instance, a maximum quantified value of 1 for a building suggests that the entire image is labelled as such. Upon closer inspection, it becomes evident that this phenomenon often occurs when the camera is near the building, resulting in the exclusion of background elements. Additionally, instances of inflated maximum pixel percentages can be attributed to misidentification, as seen with the distribution box category reaching a maximum value of 0.842. After checking, this image is found that substantial portions of sky and trees are erroneously labelled as distribution box.

Moving to the instance-unit categories, cars, bus stops, and bikes emerge as the most frequent occurrences. The mask evaluation already reveals instances where unidentified and semantic-less objects/regions are ambiguously labelled as ”bus stops” or ”distribution boxes.” The presence of instances such as a maximum of four bus stops in one image also proves the misidentification.

4.3. Examination of feature correlation

Following the established methodology workflow, the next step involves examining correlations between BE features. Initially, attention is directed towards examining the correlation matrix, illustrated in Figure 4.11. Notably, variables including buildings, sky, trees, grass, road, and sidewalks exhibit relatively high correlations.

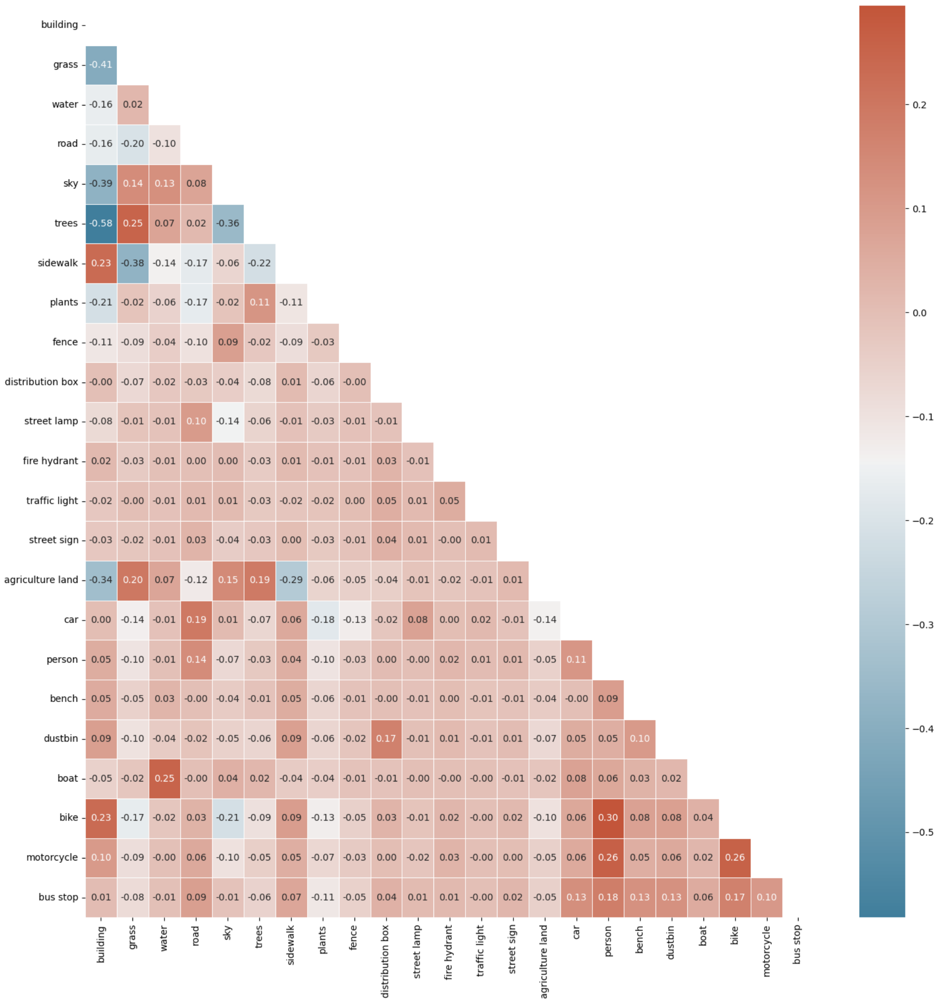


Figure 4.11: Correlation matrix

The negative correlation between buildings and trees is logical, considering they both feature prominently in vertical imagery and often overlap. An increase in tree ratio typically coincides with a reduction in visible sky proportion. This relationship holds empirically, as a broader tree canopy naturally obstructs the view of the sky. Similarly, buildings and sky exhibit a similar dynamic, with buildings frequently overlapping with the sky in upper image sections. Moreover, buildings commonly occupy vertical sections of images, thereby reducing available grass areas typically seen in lower image sections. This reasoning extends to sidewalks and grass which are often situated in lower and side image sections.

Typically, correlations exceeding 0.70 or 0.80 may indicate potential multicollinearity concerns (Belsley et al., 2005; Sabilla et al., 2019). A relevant study incorporating the micro-scale built environment features noted that even the highest correlation of 0.59 is not considered indicative of multicollinearity (Chen et al., 2024). Consequently, given that the most highly correlated value in Figure 4.11 is only

-0.58, it is reasonable to presume no significant effects on potential multicollinearity and proceed with including all variables in the subsequent analysis.

4.4. Results of multinomial logit model

Table 4.4 displays the estimated coefficients and model fits for the five models with Model 0 is the benchmark model and Model 1 to Model 4 are semantic CV-DCMs. The table also demonstrates the Value of Travel Time (VTT) [1]. In the stated choice experiment, VTT indicates the (mean) willingness to pay per month for a one-hour reduction in travel time. A VTT ranging from €217 to €231 per hour per month is reasonable, given that most respondents commute five days a week, totaling about 20 days per month. The stable values of VTT across the five models confirm the validity of the models (van Cranenburgh & Garrido-Valenzuela, 2023).

Table 4.4: Estimated coefficients and model fits for the five models

Model ID	0			1			2			3			4		
	est	r.s.e	r.t	est	r.s.e	r.t	est	r.s.e	r.t	est	r.s.e	r.t	est	r.s.e	r.t
B_Cost	0.86**	0.02	-35.60	-0.93**	0.03	-36.10	-0.93**	0.03	-36.10	-0.93**	0.03	-36.10	-0.93**	0.03	-36.10
B_Time	0.21**	0.03	-8.31	-0.24**	0.03	-9.08	-0.24**	0.03	-9.09	-0.24**	0.03	-9.22	-0.24**	0.03	-9.23
BE Features-Thing Categories															
B_DB															
B_Car															
B_Dustbin				2.32**	0.76	3.06	2.28**	0.76	3.00	2.17**	0.76	2.84	2.11**	0.76	2.76
B_Boat				1.6**	0.59	2.72	1.6**	0.59	2.70	1.75**	0.59	2.96	1.75**	0.59	2.95
B_Bike				-0.61*	0.29	-2.13	-0.63*	0.29	-2.20	-0.51	0.29	-1.75	-0.53	0.29	-1.81
B_Motorcycle				-4.27**	-4.27	-2.75	-4.22**	1.56	-2.71	-4.49**	1.57	-2.85	-4.43**	1.58	-2.81
B_BusStop				-1.98*	0.82	-2.44	-2.03*	0.82	-2.50	-2.12**	0.82	-2.59	-2.17**	0.82	-2.65
BE Features-Stuff Categories															
B_Building				1.97**	0.27	7.19	1.94**	0.28	7.08	2.18**	0.28	7.79	2.15**	0.28	7.68
B_Grass				2.08**	0.26	8.08	2.08**	0.26	8.08	2.03**	0.26	7.78	2.04**	0.26	7.81
B_Road				0.972**	0.30	3.30	0.96**	0.30	3.25	1.23**	0.30	4.07	1.21**	0.30	4.02
B_Sky				3.33**	3.33	10.10	3.3**	0.33	9.92	3.14**	0.33	9.41	3.1**	0.34	9.24
B_Trees				3.37**	0.30	11.10	3.32**	0.31	10.90	3.48**	0.31	11.30	3.43**	0.31	11.10
B_Sidewalk				1.29**	0.29	4.46	1.27**	0.29	4.38	1.48**	0.29	5.05	1.46**	0.29	4.96
B_Plants				2.98**	0.32	9.43	2.95**	0.32	9.32	2.85**	0.32	8.98	2.82**	0.32	8.87
B_Fence				0.66*	0.66	1.96	0.66	0.34	1.94	0.81*	0.34	2.39	0.81	0.34	2.37
B_Street Lamp				2.51**	0.59	4.23	2.52**	0.59	4.25	2.48**	0.59	4.20	2.49**	0.59	4.22
B_Water				3.47**	3.47	5.87	3.49**	0.59	5.91	3.91**	0.60	6.56	3.93**	0.60	6.60
B_Agriculture Land				2.63**	2.63	7.51	2.61**	0.35	7.44	2.86**	0.36	8.01	2.84**	0.36	7.97
Season1															
B_spring (3,4,5)															
B_summer (6,7,8)															
B_autumn (9,10,11)															
B_winter (12,1,2)							-0.140	0.07	-1.83				-0.15	0.08	-2.02
Low Level Features															
B_MH															
B_ML										0.54**	0.13	4.14	0.52**	0.13	3.98
B_MS															
B_PE															
B_SL															
B_SS										0.43	0.13	3.40	0.46**	0.13	3.62
B_SH															
Model Performance-Training Dataset															
Log likelihood	-5953.908			-5663.193			-5661.530			-5646.487			-5644.471		
AIC	11911.820			11362.390			11361.060			11332.970			11330.940		
BIC	11926.190			11491.780			11497.640			11476.740			11481.900		
Rho square-bar	0.122			0.162			0.162			0.164			0.165		
Model Performance-Test Datatset															
Log likelihood	-1193.654			-1161.921			-1160.800			-1156.229			-1155.079		
AIC	2391.308			2359.842			2359.599			2352.457			2352.158		
BIC	2402.457			2460.184			2465.516			2463.948			2469.224		
Number of Parameters															
	2			18			19			20			21		
Value of Travel Time [euro/hour/month]															
	216.918			228.155			228.155			231.290			231.042		

Note: p**<0.01, p*<0.05

[1] VTT = 60*(225/15) * (B_Time/B_Cost), (225/15) arises from scaling the attributes before training.

The choice dataset is divided into a training dataset and a test dataset to mitigate the risk of overfitting. The training set is utilized to train the model. The test is unseen by the model during training and used to evaluate the model's generalization performance after training. If a trained model overfits the data, disparities in performance between the training and test sets become evident. The training and test datasets consist of N = 9,784 and N = 1,948 choice observations, respectively. For more in-depth information on this dataset partitioning, please refer to the cited paper of van Cranenburgh and Garrido-Valenzuela (2023).

4.4.1. Results analysis on model fits

Table 4.4 presents a clear progression in model performance for the semantic CV-DCMs (Model 1, 2, 3, 4). A consistent trend is observed in both the training and test datasets, reflected by log-likelihood and rho-square-bar. Model 1, which comprises solely numeric attributes and BE features, shows the lowest log-likelihood and rho-square-bar, indicating poorer fit and explanatory power. However, introducing season variables in Model 2 leads to a marginal improvement, as evidenced by its slightly higher rankings. This trend continues with Model 3, incorporating LLF, achieving the second-highest values, suggesting enhanced accuracy and predictive power compared to its predecessors. It is worth mentioning that the effectiveness of seasonal variables is inferior to that of LLF in enhancing model accuracy and predictive power.

Notably, Model 4, encompassing all variables (cost, time, BE features, season variables, and LLF), consistently attains the highest scores, signifying superior performance. The increasing trend of the semantic CV-DCMs' performances highlights the importance of comprehensively incorporating diverse types of information from images for better outcomes. Furthermore, the AIC and BIC of the semantic CV-DCMs confirm the trend, aligning closely with the patterns observed in log-likelihood and rho-square-bar. However, the slight instability in the increase of AIC and BIC from Model 1 to 4 suggests nuances in model complexity.

The model performances of semantic CV-DCMs are much higher than Model 0 for any performance indicator. For instance, the log-likelihood of the test dataset of Model 0 is -1193, while the corresponding values of semantic CV-DCMs range between -1161 and -1155. Similarly, the log-likelihood of the training dataset for semantic CV-DCMs surpasses that of Model 0, as well as the rho-square-bar of the training dataset. Although an increase in parameters usually leads to higher log-likelihood and rho-square-bar for the training data, the notably higher log-likelihood of the test dataset for Model 4 compared to Model 0 underscores the semantic CV-DCM's enhanced prediction of people's choice behaviour.

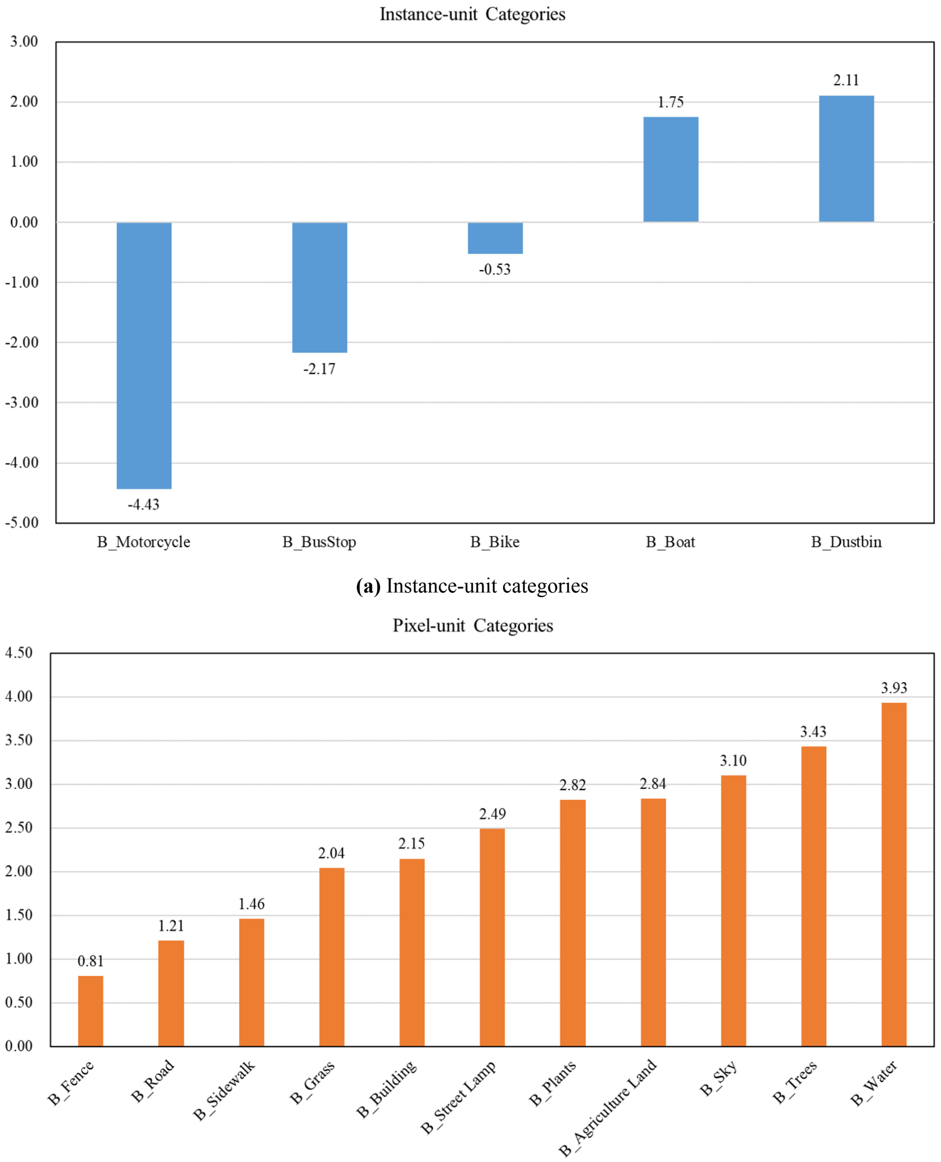
4.4.2. Results analysis on estimated parameters

Table 4.4 illustrates the estimated coefficients of the four model types. Model 4 is chosen for further analysis of coefficients as it demonstrates the highest performance based on the log-likelihood for both the training and test datasets.

It is crucial to note that the interpretation of coefficients in discrete choice models varies because of the underlying scales and units of the original variables. Although all variables in different scales have been normalized as described in 3.5.1, the coefficients represent the change in the outcome variable for a one-unit change in the predictor variable. Consequently, the magnitudes of estimated coefficients across different units of variables are not directly comparable when interpreting their impacts on choice behaviour. Given the significance of micro-scale BE features, the estimated coefficients for instance-unit and pixel-unit categories of Model 4 are visualized in Figure 4.12a and Figure 4.12b, respectively.

Other variables (housing costs, commute travel time, LLF, seasonal variables) are not shown in Figure 4.12 because they are not the primary focus of this research, and comparing their coefficients is challenging due to differing units. As shown in Table 4.4, mean lightness and standard deviation of saturation significantly enhance residential attractiveness, as expected. Additionally, winter has a negative impact on RLC, reflecting the anticipated decrease in attractiveness due to the season's association with withering landscapes. Housing costs and travel time impacts align with expectations as well.

The following discussion will address micro-scale BE features with negative and positive impacts separately, along with ranking the magnitude of these impacts.



(a) Instance-unit categories

(b) Pixel-unit categories

Figure 4.12: Visualization of estimated coefficients of model 4

Positive coefficients:
Vegetation variables, including grass, trees, and plants, are shown as positive, reasonable influences. Water and boats represent waterfront properties, offering picturesque views and recreational opportunities such as boating, fishing, or simply enjoying waterfront walks. It contributes to overall well-being and satisfaction with the living environment. Also, owning a property with waterfront access or views can be seen as a status symbol, contributing to a sense of prestige and exclusivity for residents. Dustbins also positively influence residential location choice, which dustbins near the residences can explain as representing more convenient access. The presence of the sky represents an open living environment, so it also positively influences people's choices. Fence positively impacts RLC, which is also expected since it promotes order and territory, dividing personal and open spaces (Rossetti et al., 2019; Zhanjun et al., 2022). The street lamp also positively impacts RLC, which is under expectation. The presence of street lamps can increase the safety of the residential neighbourhood.

Negative coefficients:
Both motorcycles and bikes are found to influence RLC negatively. The presence of motorcycles introduces noise, which can diminish the attractiveness of residences. Additionally, bikes in images are often observed to be stacked improperly without being placed in bicycle racks, leading to untidy piles that may detract from the residential neighbourhood's appeal. Transport-related categories, including cars, bikes, motorcycles, and bus stops, demonstrate negative influences on residents' choices. While the negative impact of bus stops might seem counterintuitive, it can be explained that most objects labelled as bus stops in images are incorrect, which are some objects not being provided semantic texts. These unknown objects may disarray the image, decreasing the corresponding residence's attractiveness. Therefore, the coefficients of bus stops will show as negative.

Ranking of magnitude of coefficients:
In Figure 4.12a, dustbins, representing convenience, have the highest impact on RLC among all instance-unit categories. Boats, which are usually together with the presence of water, have the second highest impact. This ranking suggests that people prioritize convenient facilities over waterfront views when choosing a residence. This preference indicates that practical amenities like dustbins, which simplify daily life, are more influential in residential choice than the aesthetic appeal of being near water.

Both transport-related features, bikes and motorcycles, have been noted for their negative impact on RLC. However, bus stops were omitted from the analysis due to their low identification accuracy. Notably, motorcycles stand out for their notably adverse effect on RLC, suggesting a significant decline in attractiveness where they are present. The disorder and noise associated with motorcycles can significantly detract from the appeal of residential areas.

Given that the robust standard error (r.s.t) for bikes and motorcycles is the highest among all variables (0.82 and 1.58 in Table 4.4), it is imperative to delve deeper into their impact on RLC. A high robust standard error indicates a greater degree of uncertainty or variability in estimating the respective coefficients. This prompts a pressing need for further investigation to ascertain whether motorcycles and bikes indeed exert such pronounced negative impacts on RLC, a matter of significant concern for academic researchers, urban planners, and policymakers.

Among all the pixel-unit categories in Figure 4.12b, water has the highest positive impact on RLC, indicating a strong preference for residences near canals in the Netherlands. This finding aligns with the known preference for residences surrounded by natural features. Trees, followed closely by water, further affirm this preference for residences amidst vegetation.

Interestingly, agricultural land, grass, and plants show lesser impact magnitudes than trees. As discussed in sub-section 4.2.3, these three vegetation categories are often misidentified as each other. Due to this misclassification, it becomes challenging to differentiate and rank the impacts of agricultural land, grass, and plants accurately. However, trees can be concluded to have a higher impact with greater confidence.

Street lamps and fences are facility-related categories that positively impact residential location choice. Comparing their coefficients provides insights into how people prioritize safety and order when selecting residences. Street lamps, symbolizing enhanced safety, exhibit the highest impact, reflecting residents' prioritization of safety. Fences, symbolizing order and territory, are ranked lowest regarding impact on RLC.

5 Discussion

This section delves into the limitations of employing the semantic computer vision model. First, we discuss the constraints encountered with the PSAM, which are general findings after the mask evaluation. Second, we discuss the broader limitations associated with leveraging panoptic segmentation models to quantify micro-scale BE features as inputs of choice modelling, i.e., whether the quantified features align with what people perceive in images. Subsequently, we compare and analyze the model performances of semantic CV-DCM in this thesis against the CV-DCM proposed by van Cranenburgh and Garrido-Valenzuela (2023).

5.1. Limitations of employing the semantic computer vision model

5.1.1. Limitations of the PSAM

Section 4 discussed the evaluation performance for some specific categories during the mask evaluation. There are also two key limitations concerning the PSAM. The first is related to the performance of the semantic computer vision model, which can be elaborated from three perspectives.

1) The model struggles to differentiate between similar categories, such as grass and agricultural land, as well as sidewalks and roads. Typical Dutch categories like cycling lanes also require training to make the model learn about. 2) It was observed that perfect segmentation tends to occur more frequently in simple scenes. At the same time, the model struggles to perform effectively in complex scenes featuring numerous pixel-unit categories and instances. The generated masks of images with complex scenes usually have unclear and blurred boundaries between segmented regions, and several regions are mislabeled with semantic categories. 3) The model tends to identify all of the present elements in images, even if some are not predefined. In this case, some unknown categories for the model are mistakenly recognized as predefined categories. Remarkably, by calculating total pixels and total labelled pixels, over 99% of images have more than 99% of their pixels labelled.

The above three problems could be addressed by training the model with more images of the surrounding environments of residences in the Netherlands. This additional training can enhance the model's ability to distinguish between similar categories and learn about special categories often appearing in the urban environment in the Netherlands. Furthermore, the training can equip the model to handle more complex scenes. Additionally, the trained model might have a higher threshold for labelling categories, reducing the instances where it assigns predefined categories to nearly all pixels in images. These potential solutions offer a clear path for improving the model's performance.

The second key limitation is associated with the non-overlapping property of panoptic segmentation, which aims to assign a unique label or category to each pixel in the image. This property ensures a clear and unambiguous assignment of pixels to specific objects or classes and eliminates ambiguous regions where multiple labels or categories are assigned to the same pixel. However, in instances where there is a fence with a hollow design, the model may label pixels (e.g., grass) between railings as a fence as well. The pixels between railings pose a dilemma, as they can be interpreted as both parts of the fence and grass. Deciding between these two possibilities is challenging even for humans. Importantly, this issue is not a reflection of the model's abilities but is instead tied to its inherent properties. Instances such as fences with hollow designs are infrequently shown in masks, but it is still worth mentioning this potential issue in the discussion. This highlights a limitation of the model's current design and suggests a potential area for future improvement.

5.1.2. Limitations of panoptic segmentation models for quantifying micro-scale BE features that humans perceive

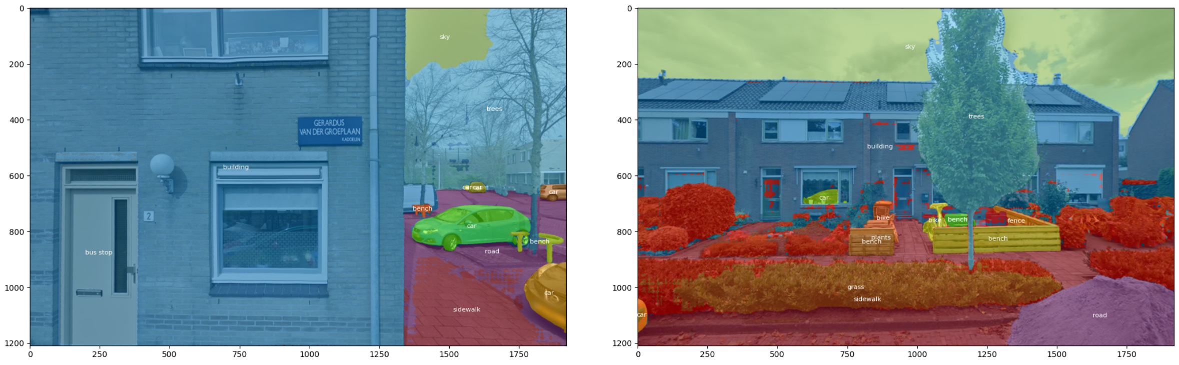
Applying panoptic segmentation models represent a need to predefine categories to be quantified in units in instances and pixels. Instance-unit and pixel-unit categories present different challenges in panoptic segmentation models regarding whether they correctly reflect how people quantify micro-scale built BE features in images. Moreover, classifying BE features into suitable semantic computer vision tasks for suitable quantified units (pixels or instances) is also a question that worth thinking.

Instance-unit categories

In traditional stated choice experiments, respondents are offered numbers or texts of alternatives and make choices. It's easy to perform the discrete choice modelling as the attributes of alternatives are in numbers and texts. However, the semantic computer vision model in this study is tasked with extracting information from images and transforming it into numbers. The key question here is whether this semantic computer vision model can accurately reproduce how people perceive the quantities of micro-scale BE features in images.

In the middle part on the right side of Figure 5.1a and the left bottom corner of Figure 5.1b, the model accurately identifies cars of petite sizes. Usually, it should be considered a high-quality mask for correctly identifying objects that are hard to be identified. However, this also questions whether these small objects affect people's decision-making. Participants may not notice these small cars when they observe these images and make choices. Taking cars occupying a small area into account may significantly influence the results of choice modelling since the segmentation unit of cars is predefined as an instance rather than a pixel. This would not be a problem for the pixel-unit categories since their segmentation results already tell the area occupied in the image.

Therefore, for instance-unit categories, it might be crucial to correctly and effectively capture instances participating in people's cognitive processes. This problem needs further research for studies that use instance segmentation or object detection tasks to quantify instances in images. Establishing a threshold of the proportion of pixels deemed ”big” enough to qualify as an instance, we can then explore whether variations in the number of instances have an impact on choice modeling outcomes.



(a) Image ID: 7

(b) Image ID: 7017

Figure 5.1: Masks that include cars in small sizes

Pixel-unit categories

The pixel-unit categories have their problems with the depth of field. It means that objects located at different depths within a scene may appear to be the same size in the image due to the effects of perspective (Forsyth & Ponce, 2002). For instance, if a semantic segmentation model identifies a building and grass in an image, it should understand that buildings typically have larger physical dimensions than grass, regardless of their apparent sizes in the image due to perspective. Figure 5.2 shows an example that grass and buildings have similar proportions of pixels identified: 0.18. However, the actual sizes of objects may contradict this visual interpretation. Human observers, equipped with contextual knowledge and semantic understanding derived from real-life experiences, can intuitively discern the actual sizes of objects despite their appearances in the image (B. Zhou et al., 2019). Semantic segmentation results of buildings and grass in Figure 5.2 are certainly accurate. Nevertheless, the accuracy may not reflect how people perceive the quantities of buildings and grass, which could affect the results of the discrete choice modelling. Incorporating depth estimation (i.e., determining the distance of objects from a camera) into semantic segmentation may be closer to how people quantified pixel-unit categories in images (Robbins et al., 2022).

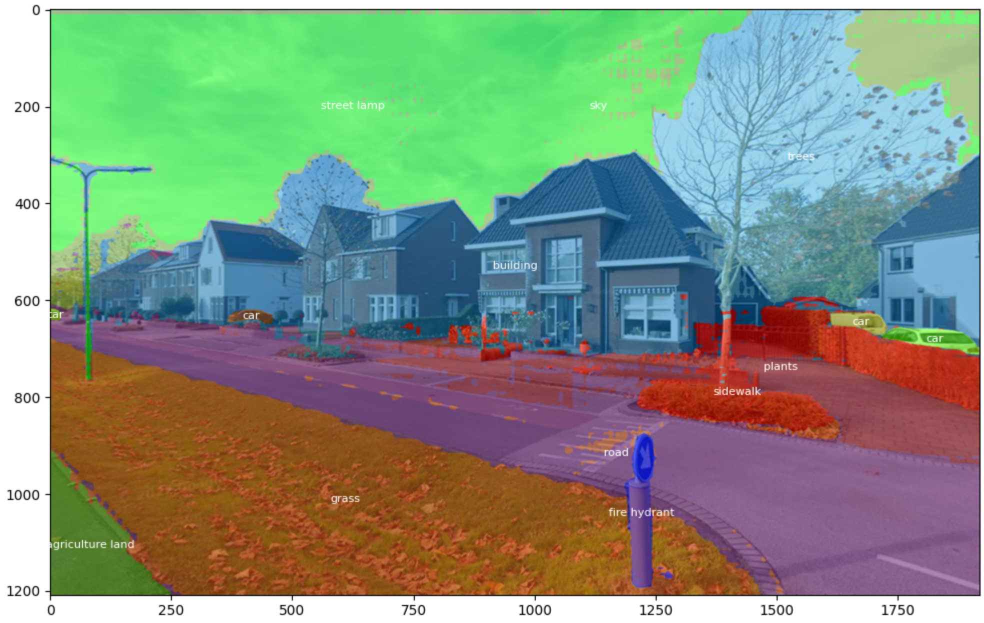


Figure 5.2: Image ID: 1020 (grass and building have the similar pixel proportions)

Classifying BE features into suitable categories

Classifying each micro-scale BE feature into a suitable category (pixel-unit or instance-unit) is a challenge. The output unit of the semantic computer vision model matters because it relates to how people perceive and process information when making choices. For instance, people will be aware of tree quantities by the proportion of trees that occupy the image rather than counting how many trees are one by one since trees represent amorphous regions. On the contrary, cars are more suitable in instance units since they are easily counted. These are two categories that are easy to classify.

However, it's important to note that even countable objects can pose challenges when they accumulate in a region of the image, as illustrated by the bikes in Figure 4.10. In such cases, the suitable unit for bikes could be a pixel, as people perceive them as amorphous regions. Similarly, street lamps and traffic lights, although currently classified into pixel-unit categories, may be better quantified in the unit of instances, as people are often more concerned with their presence or absence. These examples highlight the need for further experiments to determine the most suitable semantic computer vision tasks and output units for different BE features.

5.2. Comparison between semantic CV-DCM, lin-add RUM-MNL and CV-DCM

Table 5.1 presents a comparative analysis of the performance between the benchmark model (Model 0), the semantic CV-DCM developed in this study, and the CV-DCM proposed by van Cranenburgh and Garrido-Valenzuela (2023). Model 0 includes only cost and time variables in its utility function. Model 4, exhibiting the best performance, represents the semantic CV-DCM. Model 5 corresponds to the CV-DCM, encompassing 86 million variables, which include housing costs, travel time and image embedding that lacks interpretability. The value of time for all four models falls within a reasonable range, validating their validity (van Cranenburgh & Garrido-Valenzuela, 2023). Additionally, the three models all include housing costs and travel time, while Models 4 and 5 clearly show higher model performances than Model 0 since the two models incorporate more variables than Model 0.

Table 5.1: Comparison on model performances between semantic CV-DCM, lin-add RUM-MNL and CV-DCM

	Model Name		Semantic CV-DCM	CV-DCM
	Model ID	0	4	5
	Number of parameters	2	20	86m
Train dataset	Log likelihood	-5954	-5644.471	-5724
	Rho-square-bar	0.12	0.165	0.156
Test dataset	Log likelihood	-1194	-1155.079	-1137
Value of Time	euro/hour month	216.700	231.042	228.500

Comparing Model 4 and Model 5, while Model 4 still outperforms the training dataset, it exhibits a lower log likelihood for the test data. The test data is instrumental in evaluating a model's generalization performance after training. The lower log likelihood of the training data for Model 4 suggests that Model 5 fits the training data more closely, capturing its patterns and relationships. Conversely, the high log likelihood of the test dataset for Model 4 suggests potential overfitting of the training data. Model 5, by contrast, demonstrates superior generalization to unseen data, indicating its capability to capture underlying patterns without being overly influenced by noise or specific characteristics of the training set. Model 4's better performance on the training data may indicate its tendency to memorize the training set rather than learning the underlying patterns, resulting in poor generalization of the test data.

The comparison between Model 4 and Model 5 aligns with expectations, given that Model 5 utilizes a feature map, a deep neural network, to extract the most salient characteristics from images. With 86 million weights, the feature map encompasses a wealth of information from images, surpassing the quantified micro-scale BE features. However, Model 5's enhanced predictability comes at the cost of incorporating numerous parameters lacking behavioural meaning, rendering the analysis of choice behaviour by parameters impossible. The comparison between models 4 and 5 exemplifies the advantages and disadvantages of semantic CV-DCM and CV-DCM, respectively.

6 Conclusion

This study proposes a semantic CV-DCM that can extract information from images by a semantic computer vision model and input the information to traditional discrete model to investigate the impacts of micro-scale BE features on residential location choice behaviour. The estimated coefficients of choice modelling tell the influence of different micro-scale BE features on attractiveness of residential neighborhoods, which can provide valuable insights on how to design the appealing residential neighbborhoods in a cost-efficient way regarding the mix and quantities of micro-scale BE features.

6.1. Answers to research questions

Sub-research question 1: What micro-scale BE features are quantified by the semantic computer vision model?
The selected micro-scale BE features are buildings, grass, roads, skies, trees, sidewalks, plants, street signs, traffic lights, fences, street lamps, water, fire hydrants, distribution boxes, agricultural land, cars, motorcycles, bikes, pedestrians, benches, dustbins, boats, and bus stops.

After a review of relevant literature, it becomes evident that studies utilizing virtual tools for quantifying Built Environment (BE) features by human raters encompass a wide array of micro-scale elements. Human raters can capture intricate details within images, including diverse BE features such as litter in streets, varying levels of street lighting (car-oriented or pedestrian-oriented), and marked crosswalks. Furthermore, the assessment criteria employed in these studies exhibit high variability and subjectivity, ranging from assigning scores to sidewalk continuity (e.g., ”high” or ”low”) to binary indicators (e.g., ”1” or ”0”) denoting the presence or absence of specific features.

In contrast, studies leveraging semantic computer vision models for quantifying micro-scale BE features demonstrate a higher degree of consistency in the features they choose to investigate. These studies focus on more standardized BE features such as buildings, grass, and roads. While slight variations in the selected features may exist among different studies, the core set of features remains relatively consistent. Consequently, these studies serve as valuable references for determining which BE features to select and quantify.

Given the utilization of a zero-shot computer vision model in this study due to the absence of labelled training images, selecting frequent features from the SVI dataset is critical. Forming a category list containing the most prevalent BE features depicted in the images may decide whether the semantic computer vision model quantifies features accurately. Following a manual review of approximately

200 images, BE features appearing more than once are identified.

These identified features are then classified into appropriate semantic computer vision tasks based on their countability. Features that can be counted are categorized as instance-unit categories, such as segmentation tasks. In contrast, those that cannot be counted are designated pixel-unit categories for semantic segmentation tasks. However, specific categories, such as street lamps and traffic lights, pose challenges in classification due to occupying amorphous regions and their ambiguous countability. Street lamps, traffic lights, fire hydrants and distribution boxes are classified as pixel-unit categories.

Sub-research question 2: To which extent can the semantic computer vision model accurately quantify micro-scale BE features?
After evaluating the 400 masks and recording the incorrectly identified pixel proportions and instances for each mask, the overall segmentation outcomes were found to be satisfactory. This evaluation confirms the effectiveness of the PSAM results and holds significant implications for subsequent choice modeling procedures.

In addition to the numerical results of mask evaluation, the evaluator also documented pertinent observations about specific BE features during the assessment process. Several issues were noted. For instance, there were challenges in identifying street lamps in certain scenes. Bus stops were seldom correctly identified, but semantic-less or unknown regions often labelled as bus stops. Similar difficulties were encountered with distribution boxes. Additionally, the number of instances of bikes was frequently under-identified. Moreover, different categories of vegetation were often misidentified as one another. These findings are invaluable for gaining insight into the accuracy of quantification for various BE features.

It's evident that the zero-shot model faces limitations in accurately capturing all BE features without prior training. By having a comprehensive understanding of the accuracy of each BE feature, we can better interpret the estimated coefficients in choice modeling. On one hand, evaluation of specific categories can offer explanations if the quantified impacts of certain BE features appear unreasonable. On the other hand, ensuring that BE features are rarely misquantified enhances the reliability and validity of their impacts on Residential Location Choice (RLC). This underscores the importance of assessing the model's performance on individual BE features to bolster the credibility of choice modeling outcomes.

Sub-research question 3: How do micro-scale BE features and other attributes of residences affect people's choice behavior on residential locations?
All pixel-unit categories—water, trees, sky, agricultural land, plants, street lamps, buildings, grass, sidewalks, roads, and fences—exhibit positive coefficients in the choice modelling results. Comparing the magnitudes of these coefficients reveals the positive contribution of one pixel for each pixel-unit category to the attractiveness of residential neighbourhoods.

Among the pixel-unit categories, water and trees emerge as the most influential factors in RLC, followed by plants, agricultural land, and grass, which have similar coefficient values. However, due to challenges in the PSAM's differentiation of plants, grass, and agricultural land, it is difficult to rank their impacts conclusively.

Street lamps indicate safety and have a relatively high impact on RLC, reflecting residents' safety concerns. Fences, symbolizing order and territory, have the lowest positive impact on RLC. These two features are quantified with relatively low robust standard errors, indicating reliable measurements.

The choice modelling results show positive and negative coefficients for the instance-unit categories. Among these categories (dustbin, boat, bike, bus stop, motorcycle), dustbins represent convenience and demonstrate the highest positive impact on RLC. Adding one dustbin makes residences more attractive than adding one instance of any other instance-unit category. Boats also show relatively high impacts on RLC, indicating residents' preferences for waterfront properties.

On the other hand, bikes, bus stops, and motorcycles negatively influence RLC, reflecting that their presence makes residences less attractive. Despite these negative coefficients, determining the exact values of their negative influence on RLC is challenging. This difficulty arises from the high robust standard error associated with motorcycles and the infrequent correct identification of bus stops. In contrast, bikes may be identified accurately, but their quantization is often less than the actual instances in images.

In terms of other attributes, both the mean lightness and the standard deviation of saturation significantly enhance residential attractiveness, as expected. Brighter and more varied color saturation in an area likely makes it more visually appealing to residents. Additionally, the winter season has a negative impact on RLC, which is understandable. Images in winter typically show withering landscapes, making residences less attractive during this time. Furthermore, influences of housing costs and travel time align with expectations, as higher housing costs strain financial resources, and longer travel times reduce convenience and quality of life. Both factors make a location less desirable for potential residents.

6.2. Contributions

The contributions of this study can be divided into scientific and practical ones. The scientific contribution is as follows:

- The proposed semantic CV-DCM represents the first application of panoptic segmentation models as inputs for choice modelling. Specifically, panoptic segmentation models are utilized to appropriately categorize diverse micro-scale BE features. These models combine instance segmentation, which identifies individual objects, and semantic segmentation, which classifies each pixel into a category. This approach ensures that each micro-scale BE feature is assigned to a suitable category (i.e., instance-unit and pixel-unit categories), enhancing the accuracy and relevance of the choice modelling process.

- This thesis contributes significantly by addressing the common limitations in applying pre-trained computer vision models for choice modelling. Unlike previous studies that use these models without evaluating the accuracy of the generated masks or the overall performance on their specific datasets, this research rigorously assesses these aspects. Doing so clarifies whether the accuracies reported for pre-trained models hold when applied to new datasets. Additionally, the detailed analysis of specific categories during mask evaluation provides valuable insights into the actual accuracies of different categories, enhancing the interpretation of estimated coefficients in choice modelling. This thorough evaluation ensures a more reliable and accurate use of semantic computer vision models in understanding and predicting residential location choices.

- The proposed semantic CV-DCM, including a panoptic segmentation model, mask evaluation, and discrete choice model, is a pipeline that can be applied in other research directions. The research can be any studying contexts that require quantifying categories in images and studying their impacts on choice behaviour.

The practical contribution is as follows:

- This thesis makes a significant contribution by exploring the primarily overlooked area of RLC in relation to micro-scale BE features. While existing studies have focused on the correlation between physical and mental health and micro-scale BE features or the impact of these features on perceptions using datasets like Place Pulse 2.0, they fall short of providing comprehensive guidance on designing appealing BE near residences. This research bridges that gap by integrating the quantification of micro-scale BE features with Street View Imagery (SVI) and stated choice datasets. The findings reveal the specific impacts of these features on RLC, offering valuable insights for urban planners. By identifying the most influential elements, the study enables cost-effective restructuring of residential neighbourhoods to enhance their attractiveness and improve residents' well-being.

- This thesis provides guidance for policymakers and urban planners on enhancing residential neighbourhood attractiveness. Positive micro-scale BE features identified include water, trees, plants, agricultural land, grass, and street lamps, significantly boosting neighbourhood appeal. Water, in particular, has the highest positive impact on RLC, indicating that integrating water features such as lakes, rivers, ponds, and canals can greatly enhance neighbourhood desirability. This finding is reinforced by the positive influence of boats on RLC. Besides, trees have the strongest influence on RLC among all vegetation types, so urban planners may prioritize maintaining trees. Additionally, dustbins positively influence attractiveness, highlighting the importance of ensuring convenient waste management facilities.

- On the other hand, bikes and motorcycles have a negative impact on RLC. Policymakers should tackle these issues by strategically managing or reducing the presence of these features. For example, enhancing bike storage solutions and regulating motorcycle parking can alleviate their negative effects. By concentrating on increasing positive BE features, particularly water elements, and addressing the negative ones, urban planners can fashion more desirable and attractive residential areas, thereby enhancing the overall quality of life for residents.

6.3. Limitations and recommendations

Feature selection
Most studies utilizing semantic computer vision models for upstream tasks rely on closed-vocabulary and zero-shot (pre-trained) models. Selecting relevant categories that are aligned with the study context and frequently occurring in the local dataset is crucial. However, determining relevance and frequency can be challenging. Including too many irrelevant or infrequent categories may reduce the accuracy of quantified BE features. However, it may not affect the choice modelling results significantly. It is worth experimenting with several category lists, including different BE features, and running the semantic CV-DCM to see if the final results of choice modelling differ significantly.

Training the semantic computer vision model with local datasets
Several limitations affect the PSAM's performance in quantifying micro-scale BE features, as outlined in section 5.1.1. These limitations could be mitigated by enriching the model's training dataset with images depicting the surrounding environments of residences in the Netherlands. This additional training could improve the model's capacity to differentiate between similar categories and familiarize it with unique categories commonly found in the Dutch urban landscape. Moreover, expanded training could enable the model to handle more intricate scenes. The trained model might also adopt a higher threshold for labelling categories, reducing instances where it over-assigns predefined categories to pixels. These potential solutions provide a clear pathway for enhancing the model's performance and improving recognition, particularly for challenging categories like bus stops.

Depth of field of pixel-unit categories

Depth of field refers to the phenomenon where objects situated at varying depths within a scene may appear to be of similar size in an image, owing to perspective effects. To faithfully reproduce how individuals perceive object sizes in images, semantic segmentation algorithms must not solely account for geometric attributes but also consider semantic meanings and contextual relationships within the scene. Alternatively, integrating depth estimation (measure the distance between the object and the camera) with semantic segmentation could aid in correcting the sizes of pixel-unit categories to align more closely with human perception of objects in images.

Small objects for instance-unit categories

For instance-unit categories, particular small objects may be accurately identified by the model, although they could be too insignificant to influence human cognitive processes significantly. This issue warrants further investigation. For instance, by establishing multiple threshold values determining the proportion of pixels considered ”significant” enough to qualify as an instance, we can assess whether fluctuations in the number of instances affect choice modelling results and identify the threshold that yields optimal fit for choice modelling. Alternatively, employing semantic segmentation to measure the sizes of instance-unit categories, as demonstrated in Y. Zhang et al. (2023), presents another viable approach.

Appropriate semantic computer vision tasks for different micro-scale BE features

Determining the appropriate categorization (pixel-unit or instance-unit) for each micro-scale BE feature is critical. The choice of output unit from the semantic computer vision model is pivotal as it influences how individuals perceive and interpret information during decision-making processes. Some categories may benefit from a different unit than initially expected. For example, given the relatively small size of bikes compared to other instance-unit categories, semantic segmentation and quantifying their spatial extent might yield more precise results. Exploring which unit of measurement for categories aligns more closely with human perception of quantity is worthy of further investigation.

Heterogeneity

While this study has not accounted for individual heterogeneity, it’s essential to recognize that preferences for residential neighborhoods can vary significantly among different demographic groups (Ramírez et al., 2021). Investigating the influence of micro-scale BE features on RLC across diverse demographic segments could provide valuable insights. By understanding how different groups prioritize various neighborhood attributes, urban planners can develop more tailored policies for neighborhood design in different regions, catering to the specific preferences of predominant demographic groups.

Non-linear choice modelling

The micro-scale BE features are assumed to exert a linear additive influence on people’s choices. However, it is essential to acknowledge that this assumption may not always hold in real-world scenarios. There is a possibility that the relationships between micro-scale BE features and residential location choices are not linear additive but rather exhibit non-linear patterns. Therefore, exploring non-linear choice modelling approaches to uncover potential non-linear relationships between micro-scale BE features and choice behaviour is worthwhile. This could provide a novel understanding of how different BE features interact and influence individuals’ decisions regarding residential location.

Interactions between micro-scale BE fetures

This thesis primarily examines the individual impacts of quantified BE features, overlooking their potential interacting effects. For instance, the spatial arrangement or relationship between specific BE features within a neighbourhood could significantly influence people's choices in varying ways. Additionally, combining multiple BE features to create new variables, such as the concept of enclosure proposed by Meng et al. (2024), could be vital for understanding public sentiment and its impact on residential preferences. Therefore, future research should consider exploring the interactions between different BE features and investigating the creation of composite variables to capture more detailed neighbourhood characteristics instead of only single impacts of quantified BE features.

Find the micro-scale BE features influencing RLC by Explainable AI

The semantic CV-DCM assumes predetermined micro-scale BE features influence people's choices based on their quantified quantities. However, another approach, Explainable AI, takes a reverse stance by identifying regions in images that contribute most to the attractiveness of residences. From there, it deduces the underlying micro-scale BE features responsible for these appealing aspects. This methodology offers a complementary perspective, allowing a more complete understanding of how specific features influence residential preferences.

References

Adkins, A., Dill, J., Luhr, G., & Neal, M. (2012). Unpacking walkability: Testing the influence of urban design features on perceptions of walking environment attractiveness. *Journal of urban design*, *17*(4), 499–510. https://doi.org/10.1080/13574809.2012.706365

Arentze, T., Borgers, A., Timmermans, H., & DelMistro, R. (2003). Transport stated choice responses: Effects of task complexity, presentation format and literacy. *Transportation Research Part E: Logistics and Transportation Review*, *39*(3), 229–244. https://doi.org/10.1016/S1366-5545(02)00047-9

Ayoola, A. B., Oyetunji, A. K., Amaechi, C. V., Olukolajo, M. A., Ullah, S., & Kemiki, O. A. (2023). Determining residential location choice along the coastline in victoria island, nigeria using a factor analytical approach. *Buildings*, *13*(6), 1513. https://doi.org/10.3390/buildings13061513

Been, V., Ellen, I. G., Gedal, M., Glaeser, E., & McCabe, B. J. (2016). Preserving history or restricting development? the heterogeneous effects of historic districts on local housing markets in new york city. *Journal of Urban Economics*, *92*, 16–30. https://doi.org/10.1016/j.jue.2015.12.002

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons. https://doi.org/10.1080/00224065.1983.11978865

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

Blair, A., Ross, N. A., Gariepy, G., & Schmitz, N. (2014). How do neighborhoods affect depression outcomes? a realist review and a call for the examination of causal pathways. *Social Psychiatry and Psychiatric Epidemiology*, *49*, 873–887. https://doi.org/10.1007/s00127-013-0810-z

Cain, K. L., Millstein, R. A., Sallis, J. F., Conway, T. L., Gavand, K. A., Frank, L. D., Saelens, B. E., Geremia, C. M., Chapman, J., Adams, M. A., et al. (2014). Contribution of streetscape audits to explanation of physical activity in four age groups based on the microscale audit of pedestrian streetscapes (maps). *Social science & medicine*, *116*, 82–92. https://doi.org/10.1016/j.socscimed.2014.06.042

Cao, W., Zhou, C., Wu, Y., Ming, Z., Xu, Z., & Zhang, J. (2020). Research progress of zero-shot learning beyond computer vision. *Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part II 20*, 538–551. https://doi.org/10.1007/978-3-030-60239-0_36

Cervero, R., Sarmiento, O. L., Jacoby, E., Gomez, L. F., & Neiman, A. (2009). Influences of built environments on walking and cycling: Lessons from bogotá. *International journal of sustainable transportation*, *3*(4), 203–226. https://doi.org/10.1080/15568310802178314

Chen, M., Cai, Y., Guo, S., Sun, R., Song, Y., & Shen, X. (2024). Evaluating implied urban nature vitality in san francisco: An interdisciplinary approach combining census data, street view images, and social media analysis. *Urban Forestry & Urban Greening*, *95*, 128289. https://doi.org/10.1016/j.ufug.2024.128289

Cockx, K., & Canters, F. (2020). Determining heterogeneity of residential location preferences of households in belgium. *Applied Geography*, *124*, 102271. https://doi.org/10.1016/j.apgeog.2020.102271

Czembrowski, P., & Kronenberg, J. (2016). Hedonic pricing and different urban green space types and sizes: Insights into the discussion on valuing ecosystem services. *Landscape and Urban Planning*, *146*, 11–19. https://doi.org/10.1016/j.landurbplan.2015.10.005

Diao, M., Qin, Y., & Sing, T. F. (2016). Negative externalities of rail noise and housing values: Evidence from the cessation of railway operations in singapore. *Real Estate Economics*, *44*(4), 878–917. https://doi.org/10.1111/1540-6229.12123

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 196–212. https://doi.org/10.1007/978-3-319-46448-0_12

Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American planning association*, *76*(3), 265–294. https://doi.org/10.1080/01944361003766766

Fan, Z., Zhang, F., Loo, B. P., & Ratti, C. (2023). Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences*, *120*(27), e2220417120. https://doi.org/10.1073/pnas.2220417120

Forsyth, D. A., & Ponce, J. (2002). *Computer vision: A modern approach*. prentice hall professional technical reference. https://dl.acm.org/doi/abs/10.5555/580035

Greene, G., Fone, D., Farewell, D., Rodgers, S., Paranjothy, S., Carter, B., & White, J. (2020). Improving mental health through neighbourhood regeneration: The role of cohesion, belonging, quality and disorder. *European journal of public health*, *30*(5), 964–966. https://doi.org/10.1093/eurpub/ckz221

Guan, X., & Wang, D. (2020). The multiplicity of self-selection: What do travel attitudes influence first, residential location or work place? *Journal of Transport Geography*, *87*, 102809. https://doi.org/10.1016/j.jtrangeo.2020.102809

Hurtubia, R., Gallay, O., & Bierlaire, M. (2010). Attributes of households, locations and real estate markets for land use modelling. *SustainCity Deliverable*, *2*(1). https://www.sustaincity.ethz.ch/publications/WP_2.7_Socioeconomic_attributes.pdf

Iglesias, P., Greene, M., & Ortúzar, J. d. D. (2013). On the perception of safety in low income neighbourhoods: Using digital images in a stated choice experiment. *Choice Modelling: The State of the Art and the State of Practic*, 193–210. https://doi.org/10.4337/9781781007273.00014

Ito, K., & Biljecki, F. (2021). Assessing bikeability with street view imagery and computer vision. *Transportation research part C: emerging technologies*, *132*, 103371. https://doi.org/10.1016/j.trc.2021.103371

Jeon, J., & Woo, A. (2023). Deep learning analysis of street panorama images to evaluate the streetscape walkability of neighborhoods for subsidized families in seoul, korea. *Landscape and Urban Planning*, *230*, 104631. https://doi.org/10.1016/j.landurbplan.2022.104631

Jiang, H., Dong, L., & Qiu, B. (2022). How are macro-scale and micro-scale built environments associated with running activity? the application of strava data and deep learning in inner london. *ISPRS International Journal of Geo-Information*, *11*(10), 504. https://doi.org/10.3390/ijgi11100504

Kim, H. N., Boxall, P. C., et al. (2019). Analysis of the economic impact of water management policy on residential prices: Modifying choice set formation in a discrete house choice analysis. *Journal of choice modelling*, *33*, 100148. https://doi.org/10.1016/j.jocm.2018.07.001

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9404–9413. https://doi.org/10.48550/arXiv.1801.00868

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*. https://doi.org/10.48550/arXiv.2304.02643

Koo, B. W., Guhathakurta, S., & Botchwey, N. (2022). How are neighborhood and street-level walkability factors associated with walking behaviors? a big data approach using street view images. *Environment and Behavior*, *54*(1), 211–241. https://doi.org/10.1177/00139165211014609

Li, Y., Peng, L., Wu, C., & Zhang, J. (2022). Street view imagery (svi) in the built environment: A theoretical and systematic review. *Buildings*, *12*(8), 1167. https://doi.org/10.3390/buildings12081167

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. (2023). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*. https://doi.org/10.48550/arXiv.2303.05499

Loukaitou-Sideris, A., Liggett, R., Iseki, H., & Thurlow, W. (2001). Measuring the effects of built environment on bus stop crime. *Environment and Planning B: Planning and Design*, *28*(2), 255–280. https://doi.org/10.1068/b2642r

Lüddecke, T., & Ecker, A. (2022). Image segmentation using text and image prompts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7086–7096. https://doi.org/10.48550/arXiv.2112.10003

Ma, J., Dong, G., Chen, Y., & Zhang, W. (2018). Does satisfactory neighbourhood environment lead to a satisfying life? an investigation of the association between neighbourhood environment and life satisfaction in beijing. *Cities*, *74*, 229–239. https://doi.org/10.1016/j.cities.2017.12.008

Meng, Y., Sun, D., Lyu, M., Niu, J., & Fukuda, H. (2024). Measuring public sentiment of residential built environment through street view image and deep learning. *Available at SSRN 4725926*.

Mertens, L., Van Holle, V., De Bourdeaudhuij, I., Deforche, B., Salmon, J., Nasar, J., Van de Weghe, N., Van Dyck, D., & Van Cauwenberg, J. (2014). The effect of changing micro-scale physical environmental factors on an environment's invitingness for transportation cycling in adults: An exploratory study using manipulated photographs. *International journal of behavioral nutrition and physical activity*, *11*, 1–12. https://doi.org/10.1186/s12966-014-0088-x

Molina-García, J., Campos, S., García-Massó, X., Herrador-Colmenero, M., Gálvez-Fernández, P., Molina-Soberanes, D., Queralt, A., & Chillón, P. (2020). Different neighborhood walkability indexes for active commuting to school are necessary for urban and rural children and adolescents. *International journal of behavioral nutrition and physical activity*, *17*, 1–11. https://doi.org/10.1186/s12966-020-01028-0

Moniruzzaman, M., & Páez, A. (2012). A model-based approach to select case sites for walkability audits. *Health & Place*, *18*(6), 1323–1334. https://doi.org/10.1016/j.healthplace.2012.09.013

Nguyen, Q. C., Belnap, T., Dwivedi, P., Deligani, A. H. N., Kumar, A., Li, D., Whitaker, R., Keralis, J., Mane, H., Yue, X., et al. (2022). Google street view images as predictors of patient health outcomes, 2017–2019. *Big data and cognitive computing*, *6*(1), 15. https://doi.org/10.3390/bdcc6010015

Perkins, D. D., Meeks, J. W., & Taylor, R. B. (1992). The physical environment of street blocks and resident perceptions of crime and disorder: Implications for theory and measurement. *Journal of environmental psychology*, *12*(1), 21–34. https://doi.org/10.1016/S0272-4944(05)80294-4

Pfeiffer, D., & Cloutier, S. (2016). Planning for happy neighborhoods. *Journal of the American planning association*, *82*(3), 267–279. https://doi.org/10.1080/01944363.2016.1166347

Poudel, N., & Singleton, P. A. (2022). Preferences for roundabout attributes among us bicyclists: A discrete choice experiment. *Transportation research part A: policy and practice*, *155*, 316–329. https://doi.org/10.1016/j.tra.2021.11.023

Qi, M., Dixit, K., Marshall, J. D., Zhang, W., & Hankey, S. (2022). National land use regression model for no2 using street view imagery and satellite observations. *Environmental Science & Technology*, *56*(18), 13499–13509. https://doi.org/10.1021/acs.est.2c03581

Ramírez, T., Hurtubia, R., Lobel, H., & Rossetti, T. (2021). Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety. *Landscape and Urban Planning*, *208*, 104002. https://doi.org/10.1016/j.landurbplan.2020.104002

Robbins, E., Dubbelde, D., Wegner-Clemens, K., & Shomstein, S. (2022). Contextual effects on size perception of semantic objects. *Journal of Vision*, *22*(14), 4254–4254. https://doi.org/10.1167/jov.22.14.4254

Rossetti, T., Guevara, C. A., Galilea, P., & Hurtubia, R. (2018). Modeling safety as a perceptual latent variable to assess cycling infrastructure. *Transportation Research Part A: Policy and Practice*, *111*, 252–265. https://doi.org/10.1016/j.tra.2018.03.019

Rossetti, T., Lobel, H., Rocco, V., & Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. *Landscape and urban planning*, *181*, 169–178. https://doi.org/10.1016/j.landurbplan.2018.09.020

Sabilla, S. I., Sarno, R., & Triyana, K. (2019). Optimizing threshold using pearson correlation for selecting features of electronic nose signals. *Int. J. Intell. Eng. Syst*, *12*(6), 81–90. https://doi.org/10.22266/ijies2019.1231.08

Sallis, J. F., Carlson, J. A., Ortega, A., Allison, M. A., Geremia, C. M., Sotres-Alvarez, D., Jankowska, M. M., Mooney, S. J., Chambers, E. C., Hanna, D. B., et al. (2022). Microscale pedestrian streetscapes and physical activity in hispanic/latino adults: Results from hchs/sol. *Health & place*, *77*, 102857. https://doi.org/10.1016/j.healthplace.2022.102857

Schirmer, P. M., Van Eggermond, M. A., & Axhausen, K. W. (2014). The role of location in residential location choice models: A review of literature. *Journal of Transport and Land Use*, *7*(2), 3–21. https://www.jstor.org/stable/26202678

Steinmetz-Wood, M., El-Geneidy, A., & Ross, N. A. (2020). Moving to policy-amenable options for built environment research: The role of micro-scale neighborhood environment in promoting walking. *Health & place*, *66*, 102462. https://doi.org/10.1016/j.healthplace.2020.102462

Steinmetz-Wood, M., Velauthapillai, K., O'Brien, G., & Ross, N. A. (2019). Assessing the micro-scale environment using google street view: The virtual systematic tool for evaluating pedestrian streetscapes (virtual-steps). *BMC public health*, *19*, 1–11. https://doi.org/10.1186/s12889-019-7460-3

Tirachini, A., Hurtubia, R., Dekker, T., & Daziano, R. A. (2017). Estimation of crowding discomfort in public transport: Results from santiago de chile. *Transportation Research Part A: Policy and Practice*, *103*, 311–326. https://doi.org/10.1016/j.tra.2017.06.008

van Cranenburgh, S., & Garrido-Valenzuela, F. (2023). Computer vision-enriched discrete choice models, with an application to residential location choice. *arXiv*. https://doi.org/10.48550/arXiv.2308.08276

Wang, K., & Ozbilen, B. (2020). Synergistic and threshold effects of telework and residential location choice on travel time allocation. *Sustainable Cities and Society*, *63*, 102468. https://doi.org/10.1016/j.scs.2020.102468

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

Wilson, J. Q., & Kelling, G. L. (2017). The police and neighborhood safety broken windows. In *Social, ecological and environmental theories of crime* (pp. 169–178). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315087863-11/police-neighborhood-safety-broken-windows-james-wilson-george-kelling

Wu, W., Ma, Z., Guo, J., Niu, X., & Zhao, K. (2022). Evaluating the effects of built environment on street vitality at the city level: An empirical research based on spatial panel durbin model. *International Journal of Environmental Research and Public Health*, *19*(3), 1664. https://doi.org/10.3390/ijerph19031664

Wu, Y.-T., Nash, P., Barnes, L. E., Minett, T., Matthews, F. E., Jones, A., & Brayne, C. (2014). Assessing environmental features related to mental health: A reliability study of visual streetscape images. *BMC Public Health*, *14*, 1–10. https://doi.org/10.1186/1471-2458-14-1094

Xu, J., Liu, Y., Liu, Y., An, R., & Tong, Z. (2023). Integrating street view images and deep learning to explore the association between human perceptions of the built environment and cardiovascular disease in older adults. *Social Science & Medicine*, *338*, 116304. https://doi.org/10.1016/j.socscimed.2023.116304

Yassin, A. M., Diah, M. L. M., Safian, E. E. M., Yahya, M. Y., Mohammad, S., & Cheng, L. S. (2019). Determining the impact of aircraft noise towards residential property price. *MATEC Web of Conferences*, *266*, 02005. https://doi.org/10.1051/matecconf/2019266 02005

Yue, X., Antonietti, A., Alirezaei, M., Tasdizen, T., Li, D., Nguyen, L., Mane, H., Sun, A., Hu, M., Whitaker, R. T., et al. (2022). Using convolutional neural networks to derive neighborhood built environments from google street view images and examine their associations with health outcomes. *International Journal of Environmental Research and Public Health*, *19*(19), 12095. https://doi.org/10.3390/ijerph191912095

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.2018.08.020

Zhang, L., Han, X., Wu, J., & Wang, L. (2023). Mechanisms influencing the factors of urban built environments and coronavirus disease 2019 at macroscopic and microscopic scales: The role of cities. *Frontiers in public health*, *11*, 1137489. https://doi.org/10.3389/fpubh.2023.1137489

Zhang, Y., Zhao, H., Li, Y., Long, Y., & Liang, W. (2023). Predicting highly dynamic traffic noise using rotating mobile monitoring and machine learning method. *Environmental research*, *229*, 115896. https://doi.org/10.1016/j.envres.2023.115896

Zhanjun, H., Wang, Z., Xie, Z., Wu, L., & Chen, Z. (2022). Multiscale analysis of the influence of street built environment on crime occurrence using street-view images. *Computers, Environment and Urban Systems*, *97*, 101865. https://doi.org/10.1016/j.compenvurbsys.2022.101865

Zhao, Y., Wang, S., Chen, D., Huang, K., Zhang, S., Qiu, W., & Li, W. (2023). Estimating the impacts of seasonal variations of streetscape on dockless bike sharing trip with street view images and computer vision. *The International Conference on Computational Design and Robotic Fabrication*, 211–224. https://doi.org/10.1007/978-981-99-8405-3_18

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, *127*, 302–321. https://doi.org/10.1007/s11263-018-1140-0

Zhou, H., Liu, L., Lan, M., Zhu, W., Song, G., Jing, F., Zhong, Y., Su, Z., & Gu, X. (2021). Using google street view imagery to capture micro built environment characteristics in drug places, compared with street robbery. *Computers, Environment and Urban Systems*, *88*, 101631. https://doi.org/10.1016/j.compenvurbsys.2021.101631