# Pixels · People · Places

## Computer Vision and Image Embeddings for Perception-Aware Urban Analytics

*Francisco Garrido-Valenzuela*

# Pixels · People · Places

Computer Vision and Image Embeddings for Perception-Aware Urban Analytics

# Pixels · People · Places

Computer Vision and Image Embeddings for
Perception-Aware Urban Analytics

## Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus Prof. dr. ir. H. Bijl

Chair of the Board for Doctorates
to be defended publicly on
Wednesday 22, April 2026 at 17:30

by

**Francisco Orlando Garrido Valenzuela**

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

| | |
|---|---|
| Rector Magnificus | Chairperson |
| Prof. dr. O. Cats | Delft University of Technology, promotor |
| Dr. ir. S. van Cranenburgh | Delft University of Technology, promotor |

Independent members:

| | |
|---|---|
| Prof. dr. ir. A. Bozzon | Delft University of Technology |
| Prof. dr. M. Hebart | Max Planck Institute, Germany |
| Dr. F. Duarte | Massachusetts Institute of Technology, USA |
| Dr. C.M. Lima Azevedo | Technical University of Denmark, Denmark |
| Prof. dr. ir. L.A. Tavasszy | Delft University of Technology, reserve member |


The research in this thesis is supported by the TU Delft AI Labs programme. In CityAI Lab, data, AI, and behavioral theory come together for the development of more attractive and livable cities.

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

*Front & Back:* Beautiful cover art that captures the entire thesis in a single illustration.

**TRAIL Thesis Series no. T2026/10, The Netherlands Research School TRAIL**

TRAIL
P.O. BOX 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-385-5



Printed in the Netherlands

*To my grandparents, parents,*
*brother and Conagu.*

# Contents

# Preface

This dissertation is the result of five years of research conducted within the CityAI Lab at Delft University of Technology (TU Delft). This Lab is part of the broader AI Initiative of TU Delft, a university-wide effort supported by national funding to expand interdisciplinary research and education in Artificial Intelligence (AI). Within this initiative, multiple laboratories explore how AI can be applied across diverse domains, from robotics and computer science to public policy, architecture, and urban planning.

The CityAI Lab brings together researchers interested in understanding cities through the lens of data, behavior, and computational methods. Its work spans topics such as urban segregation, road traffic interactions, noise perception, and human behavior in urban environments. The research presented in this dissertation contributes to this broader agenda by investigating how human perception of urban space can be incorporated into analytical models of cities using computer vision and machine learning techniques.

## A path of constant exploration

If there is one common thread that explains why I am sitting here (the Netherlands) today writing this preface, it is a restless curiosity and a strong resistance to being confined to a rigid professional path. I have always been drawn to change and diversity of thought.

This tendency was already present when I finished secondary school. Choosing a field of study at university was a very challenging task. I debated whether to pursue astronomy, engineering, or medicine, as all seemed feasible and fascinating. Ultimately, I chose engineering. This decision was made to keep my options broad, leaving many doors open while building a strong analytical foundation.

Once within engineering, I faced the next important decision: choosing my specialization. My interests gravitated toward computer science, mining engineering, and transportation. This time, I also wanted my work to have a tangible societal impact. Therefore, I chose transportation engineering. It emerged as the perfect bridge. Transportation engineering is a deeply quantitative discipline that directly shapes how people live and experience cities, while also offering abundant opportunities to work with coding and technology.

## From practice back to academia

I pursued both my undergraduate and master's degrees in transportation engineering at the Pontificia Universidad Católica (PUC) de Chile. However, finishing my studies was not the end of my exploration. I did not want to settle into a single, predictable lane. I wanted to see how transportation and urban systems work from every possible angle. I wanted to try it all.

This drive took me first to the public sector, working as an advisor for the Chilean Ministry of Transport. Specifically, working alongside an amazing team in the public transport directorate, I saw firsthand how data collides with the complex reality of public policy. Then, to experience the other side of the same coin, I pivoted to the private sector. I joined a public transport dispatching startup to experience the (private) operational side of technology on mobility. Eventually, an opportunity to teach programming courses (for transport engineers) and conduct research brought me full circle, back into the academic sphere at PUC.

None of these career moves were part of a grand master plan. They were deliberate choices driven by my need to keep exploring, adapting, learning, and tackling problems through new lenses. It was exactly this dynamic mindset that finally materialized into the decision to pursue a PhD. TU Delft offered the perfect next step: an environment that actively encouraged bridging transportation engineering and urban behaviors with artificial intelligence.

## Connecting the dots at TU Delft

This constant need to look beyond a single field deeply shaped the core of this dissertation. When I set out to study how people perceive urban spaces at scale, applying computer vision and machine learning was the initial starting point (as I am in the City**AI** lab).

However, consistent with my previous decisions, I wanted to go further. I quickly realized that relying solely on artificial intelligence was not enough to truly capture the nuances of human perception. The research demanded a broader perspective. Once again, I decided to cross disciplinary lines, this time incorporating psychological models into my analytical frameworks. This combination allowed the work to go beyond a purely technical application, touching on the fundamental human experience of the city.

Looking back, the existence of this thesis is simply the latest milestone in an ongoing journey of decisions, pivots, and continuous exploration. It reflects more than just five years of academic work. It represents a lifelong preference for exploring intersections rather than following straight lines.

*Francisco Orlando Garrido Valenzuela*
*Delft, April 2026*

# Summary

Artificial intelligence, especially computer vision (CV), is reshaping how cities are studied and designed. Street-level imagery (SLI) carries multiple layers of urban information: infrastructure, design, vegetation, human activity, and beyond. Moreover, when paired with human input, these images also reveal how places are perceived. Over the past decade, many methods have either extracted physical components from images or predicted perceptions from those components. What remains uncommon, however, is a theory-guided, reproducible framework that coherently integrates both layers. Without such a framework, studies tend to describe what cities contain without explaining how they feel, which limits attribution of perceptions to specific components, the transfer of insights across cities, and the inclusion of subjective dimensions in public decision-making. Here, an integrated framework means a workflow that (i) defines what images encode in terms of components and perceptual conditions; (ii) specifies procedures to extract each layer at city scale; (iii) identifies when and how to include human-in-the-loop feedback to safeguard perceptual validity and local context meaning; and (iv) links both layers to interpretable behavioral models that attribute effects to concrete components. This thesis develops, operationalizes, and demonstrates such a framework connecting pixels, people, and places.

The research unfolds through six interrelated studies. First, an *image typology* distinguishes physical components from perceptual conditions, providing a common vocabulary and operational criteria for image-based urban research. Two subsequent studies build models for large-scale component extraction: *Where Are the People?* assembles a pipeline to detect people and street elements in millions of images and relates them to morphological indicators; *Street Embeddings* learns transferable visual representations that recover functional and morphological street typologies without intensive labeling. To connect the physical and the perceptual dimensions, *PixelSurvey* offers a modular, open-source platform for image-based surveys (stated choice, similarity judgments, and Likert scales), standardizing stimulus control, randomization, and data export. Using these data, *From Pixels to Perceptions* trains a supervised embedding model with human similarity judgments to align visual representations with perceptual structure. Finally, *Computer-Vision–Enriched Discrete Choice Models (CV–DCM)* integrates image embeddings into random-utility models, linking visual attributes and perceptions to choices in an interpretable manner.

Taken together, the thesis shows that (a) SLI can be turned into structured data of urban form about components and conditions; (b) learned spatial representations recover meaningful, transferable typologies; (c) locally sourced, human-in-the-loop supervision improves the perceptual relevance of spatial embeddings; and (d) behavioral models can incorporate visual information to anticipate how the built environment may

influence perceptions, preferences, and choices. This, thereby enables ex ante appraisal of functional and experiential impacts. The work also offers policy guidance: use image-based surveys to broaden participation, strengthen governance around visual data, and provide a practical pathway for incorporating urban perceptions into the appraisal of urban-renewal projects.

Limitations and future directions are clear. Images can introduce important biases. For instance, uneven spatial coverage and licensing rules limit transferability. Also, latent urban representations (embeddings) remain difficult to interpret, calling for more transparent models that clarify what AI captures. Regarding perceptions, two caveats are central: (i) perception measures derived from images are local and cultural in scope and should not be generalized uncritically, and (ii) quantification of perceptions are proxies of lived experience, not the experience itself. Acknowledging these limits and reinforcing governance, the thesis charts a path toward perception-aware urban analytics that is scientifically rigorous and socially useful.

# Samenvatting

*Artificial Intelligence* (AI), en in het bijzonder *computer vision* (CV), verandert ingrijpend hoe we steden bestuderen en ontwerpen. Beelden op straatniveau (*Street-level imagery*, SLI) bevatten meerdere lagen met informatie over de stad, onder andere over infrastructuur, ontwerp, begroeiing/vegetatie en menselijke activiteit. In combinatie met informatie over de perceptie van mensen maken deze beelden ook inzichtelijk hoe plekken worden waargenomen. In het afgelopen decennium zijn tal van methoden ontwikkeld die ofwel tastbare componenten uit beelden extraheren, ofwel percepties op basis van die componenten voorspellen. Wat echter nog zelden voorkomt, is een op theorie gebaseerde, reproduceerbare kader die beide lagen op een coherente manier integreert. Zonder zo'n kader beschrijven studies vaak wát steden bevatten, zonder te verklaren hoe ze aanvoelen; daardoor blijft onduidelijk welke componenten specifieke percepties oproepen, is vergelijkbaarheid van inzichten tussen steden beperkt en is het lastig om de subjectieve dimensie in beleid mee te nemen. Een geïntegreerd kader houdt een werkwijze in die (i) beschrijft welke componenten en perceptuele condities de beelden bevatten; (ii) procedures specificeert om beide lagen op stedelijke schaal te extraheren; (iii) aangeeft wanneer en hoe *human-in-the-loop* (mens-in-de-lus) nodig is om perceptuele validiteit en contextafhankelijke betekenis te borgen; en (iv) beide lagen koppelt aan interpreteerbare gedrags-/keuzemodellen die effecten toeschrijven aan concrete componenten. Deze dissertatie ontwikkelt, operationaliseert en demonstreert zo'n kader en verbindt daarmee pixels, mensen en plekken.

De studie bestaat uit zes onderling verbonden onderzoeken. Eerst introduceer ik een *Image typology* die fysieke componenten onderscheidt van perceptuele condities, Deze image typology biedt zo een gedeeld vocabulaire en operationele criteria voor beeldgestuurde stedelijke analyse. Vervolgens ontwikkelen twee studies modellen voor grootschalige componentextractie: *Where Are the People?* bouwt een pijplijn om personen en straatelementen in miljoenen beelden te detecteren en relateert deze aan morfologische indicatoren; *Street Embeddings* leert overdraagbare visuele representaties die functionele en morfologische straattypologieën reconstrueren zonder expliciete annotatie. Om het fysieke en het perceptuele te verbinden, biedt *PixelSurvey* een modulair, open-source platform voor beeldgebaseerde enquêtes (*stated choice*, gelijkenisbeoordeling, Likert-schalen), met gestandaardiseerde stimuluscontrole, randomisatie en data-export. Op basis hiervan traint *From Pixels to Perceptions* een *supervised* embedding-model met menselijke gelijkenisbeoordeling, zodat visuele representaties aansluiten bij de perceptuele structuur. Tenslotte, integreert *Computer-Vision–Enriched Discrete Choice Models (CV–DCM)* beeld-embeddings in *random-utility* modellen, waardoor visuele kenmerken en percepties op een interpreteerbare manier aan keuzes worden gekoppeld.

Gezamenlijk laten de studies zien dat (a) SLI kan worden omgezet in gestructureerd data over stedelijke vorm; (b) geleerde representaties betekenisvolle en overdraagbare typologieën opleveren; (c) (contextafhankelijke) human-in-the-loop-supervisie de perceptuele relevantie van ruimtelijke embeddings versterkt; en (d) gedragsmodellen visuele informatie kunnen opnemen om te anticiperen hoe de gebouwde omgeving percepties, voorkeuren en keuzes kan beïnvloeden—en daarmee ex ante-beoordeling van functionele en ervaringsgerichte impact mogelijk maken. Daarmee biedt dit proefschrift praktisch advies voor beleid: zet beeldgebaseerde enquêtes in om participatie te verbreden, ontwikkel beleid rondom visuele input, en betrek stadspercepties bij de beoordeling van stedelijke vernieuwingsprojecten.

Tegelijk zijn de beperkingen en toekomstige onderzoeksrichtingen helder. Beelden kunnen belangrijke bias introduceren; ongelijke ruimtelijke dekking en licentie-/eigendomsregels beperken de overdraagbaarheid; en latente representaties (embeddings) blijven lastig te interpreteren, wat vraagt om transparantere modellen die expliciteren wat AI vastlegt. Voor percepties gelden twee kernpunten: metingen op basis van beelden (i) bevatten een lokale en culturele context en vragen om terughoudend generaliseren, en (ii) kwantificatie van perceptie blijft slechts een benadering van de menselijke ervaring, ze vervangen de ervaring niet. Door deze grenzen expliciet te erkennen en governance te versterken, schetst de dissertatie een pad naar perceptie-bewuste stedelijke analytics die wetenschappelijk robuust én maatschappelijk bruikbaar is.

# Resumen

La inteligencia artificial (IA), y en particular la visión por computador (CV), está transformando la manera en que estudiamos y diseñamos las ciudades. Las imágenes a nivel de calle (*street-level imagery*, SLI) contienen múltiples capas de información urbana: infraestructura, diseño, vegetación y también actividad humana. Además, cuando se combinan con la participación de las personas, facilitan la identificación de cómo se perciben los lugares. En la última década han surgido numerosos métodos que extraen componentes tangibles de las imágenes o que predicen percepciones a partir de esos componentes. Aun así, es poco común encontrar un marco integrado, guiado por teoría y reproducible, que analice ambas capas de información de forma coherente. Sin ese marco, los estudios suelen describir qué contienen las ciudades, independientemente de cómo se sienten. Por ello, les cuesta explicar qué componentes generan determinadas percepciones, transferir resultados entre ciudades, o sustentar decisiones públicas incorporando la dimensión subjetiva. Por "marco integrado" entendemos un flujo de trabajo que: (i) define qué información pueden entregar las imágenes en términos de componentes y condiciones perceptuales; (ii) establece procedimientos para extraer cada capa a escala urbana; (iii) especifica cuándo y cómo incorporar retroalimentación de personas (human-in-the-loop) en el modelado para capturar el valor perceptual y el significado local; y (iv) conecta ambas capas con modelos de comportamiento interpretables que permitan atribuir efectos a componentes concretos. Esta tesis desarrolla, operacionaliza y demuestra ese marco, vinculando píxeles, personas y lugares.

La investigación se articula en seis estudios interrelacionados. Primero, una *tipología de imágenes* distingue componentes tangibles de condiciones perceptuales. Esta tipología aporta un vocabulario común y criterios operativos para el uso de imágenes en análisis urbano. Luego, dos estudios desarrollan modelos para la extracción de componentes a gran escala: *Where Are the People?* crea un *pipeline* para detectar personas y elementos urbanos en millones de imágenes y los relaciona con indicadores morfológicos; *Street Embeddings* genera representaciones visuales transferibles (*embeddings*) que capturan tipologías funcionales y morfológicas de calles sin necesidad de etiquetado explícito. Con el objetivo de conectar lo físico con lo perceptual, *PixelSurvey* ofrece una plataforma modular y de acceso abierto para implementar encuestas con imágenes (elección declarada, juicios de similitud, escalas tipo Likert), estandarizando el control de estímulos, la aleatorización y la exportación de datos. Con esos insumos, *From Pixels to Perceptions* entrena un modelo de *embeddings* supervisado con juicios de similitud humanos para alinear las representaciones visuales con la estructura perceptual. Finalmente, *Computer-Vision–Enriched Discrete Choice Models (CV–DCM)* integra *embeddings* de imágenes en modelos de utilidad aleatoria, conectando atributos visuales y percepciones con elecciones de forma interpretable.

En conjunto, la tesis muestra que: (a) las SLI pueden convertirse en evidencia estructurada sobre la forma urbana; (b) las representaciones urbanas aprendidas recuperan tipologías significativas y transferibles; (c) la supervisión local con participación humana mejora la relevancia perceptual de los embeddings urbanos/espaciales; y (d) los modelos de comportamiento pueden incorporar información visual para anticipar cómo el entorno construido podrían influir en percepciones, preferencias y decisiones. Con ello, se habilita la evaluación *ex ante* de impactos fisicos/funcionales y experienciales. Además, se proponen orientaciones de política pública: usar encuestas con imágenes para ampliar la participación y fortalecer la gobernanza mediante datos visuales, y proveer un método que permita incorporar el uso de percepciones urbanas en la evaluación de proyectos de renovacion urbana.

Las limitaciones y líneas futuras son claras. A pesar del valor informativo que las imágenes aportan, pueden introducir sesgos relevantes. Por ello, es clave reconocer que la cobertura espacial desigual y las restricciones de propiedad y licenciamiento limitan la transferibilidad. Por otro lado, las representaciones latentes urbanos (*embeddings*) siguen teniendo una baja interpretabilidad. Es importante desarrollar modelos más transparentes que permitan entender lo que los modelos de IA capturan. Respecto a las percepciones, conviene distinguir que las percepciones medidas a partir de imágenes: (i) tienen un alcance local y cultural, por lo que no deben generalizarse de forma directa; y (ii) son una aproximación de la experiencia vivida, no la experciencia misma. Reconociendo estos límites y reforzando la gobernanza, esta tesis traza un camino hacia una analítica urbana sensible a la percepción, rigurosa en lo científico y socialmente útil.

# Chapter 1

# Introduction

Artificial Intelligence (AI) is reshaping how we study the world—but its value depends on how we critique it, adapt it, and apply it. This thesis explores how AI can be directed toward richer and more insightful forms of urban analysis. This enables a systematic approach to capturing urban form and developing deeper, human-centered insights into how cities are perceived.

## 1.1 Motivation

AI lies at the core of a new industrial revolution that is transforming how we generate and use information—opening new possibilities for understanding and designing cities. Over the past decade, advances in machine learning models, data availability, and computational power have enabled considerable progress across the entire field of AI. This includes image interpretation [1, 2], natural language understanding [3, 4], generative modeling [5], and predictive analysis [6]. In urban analytics, these advances offer new ways to observe, model, and interpret people and cities. Urban analytics refers to urban research that leverages new data sources, such as sensors, social media, and map platforms, among others [7]. These new advances facilitate capturing patterns, behaviors, and forms of urban life that are difficult to reach with traditional data sources or analytical frameworks.

Among the many developments within AI and supported by the growing availability of image data, computer vision (CV) has become a particularly powerful tool for urban analysis. Images are exceptionally rich carriers of urban information: they encode physical attributes such as buildings, vegetation, infrastructure, and people, as well as perceptual cues such as aesthetics, spaciousness, or disorder. Importantly, images are also intuitively understandable for people, which makes them an accessible medium for participatory and perception-based research [8]. Additionally, the widespread use of social media and mapping platforms [9–11], along with the ubiquity of cameras in smartphones, vehicles, and surveillance systems, has made street-level imagery abundant and continuously updated. Cities are now photographed from multiple angles and at different times, creating vast visual archives of urban life. This abundance enables large-scale, fine-grained analyses that were previously unfeasible with traditional data sources. Modern CV models can detect and classify objects [12], segment features [13],

track objects [14], classify scenes [15], and embed whole images into high-dimensional spaces that preserve semantic similarity [16, 17]. In other words, raw pixels can be transformed into structured data that supports empirical and model-driven studies.

Despite the growing penetration of imagery and advances in CV tools in urban studies, theory-guided workflows that integrate physical and perceptual layers about urban spaces remain scarce. Clear guidance is still missing on what type of urban information images encode, how to extract that information systematically, and when and how to incorporate human-in-the-loop steps to preserve local perceptual insight. Much of the existing CV-based research either focuses on components—physical attributes such as buildings, vegetation, or people—or on conditions—perceptual qualities such as safety, beauty, or order [18]. These two perspectives, one grounded in physical structure and the other in human experience, are rarely connected at scale within a coherent, theory-driven workflow. This separation is problematic because the visual form and perceptual feel of streets jointly influence walking behavior, safety outcomes, housing choices, and well-being [19, 20]. Research efforts often operate in isolation, with little integration between the multiple "layers" of urban systems (i.e., components, conditions, and behaviors), and limited capacity to model their interactions [7]. Moreover, the field lacks a widely adopted typology of the kinds of information urban images encode. Such a typology would enable the scalable extraction of both physical and perceptual attributes and indicate when human-in-the-loop input is needed to properly incorporate (local) residents' preferences and perceptions. As a result, urban research has yet to capitalize on imagery's potential as a comprehensive source of information. Computer-vision-based planning tools remain fragmented and only offer partial views of how urban environments are structured and experienced.

This thesis responds to the lack of precise definitions of what urban images encode and an integrated framework for working with imagery in urban analytics. It develops methods to extract both components and conditions, establishes a conceptual typology to distinguish them, and introduces tools for collecting perceptual judgments at scale. Together, these contributions lay the groundwork for analyzing how physical structure and perceptual experience can be studied in parallel, and for combining them within models that link urban form with human experience.

## 1.2 Research background

This section examines images and computer vision in urban analytics. It begins by positioning images within the urban research field, then discusses images as a data modality, and introduces computer vision methods and image embeddings. It next reviews how CV approaches have been applied to analyze urban space and to infer preferences and perceptions. Finally, it highlights the absence of an integrated framework that connects the physical structure of urban space with its more human-interpretable qualities.

### 1.2.1 Images in the urban field

Urban research has entered an image era. Images are among the most information-dense and intuitively interpretable data formats available. Every day, billions of photographs are captured by smartphones, social media users, mapping platforms, vehicles, and surveillance systems [21]. These images document the urban environment from multiple angles, at different times, and in high spatial resolution. As a result, imagery has become a ubiquitous and information-rich source for urban analysis. In particular, street-level images, which are photographs taken at street level in urban areas, can reveal a wide range of urban attributes—such as built form, vegetation, weather conditions, transportation modes, and the presence of people [7]. At the same time, imagery can evoke perceptions of spaciousness, cleanliness, beauty, or safety [22, 23]. Understanding this dual capacity of imagery is relevant for designing tools that bridge spatial analysis with lived experience.

To illustrate this, consider the pair of street-level images shown in Figure 1.1, both taken from nearly the same location in Delft, the Netherlands, but separated by 14 years. The first image, captured in 2008, shows the area before a major infrastructural intervention in which the railway was moved underground. The main elements visible are the concrete-elevated railway, parked cars, and leafless winter trees. The second image, taken more recently, reveals the transformed streetscape: a water feature, green trees, and a clear blue sky now dominate the scene. Although both images depict the same location, the visual composition and the feelings it evokes differ. The 2008 scene (a) may convey a sense of heaviness, enclosure, or insecurity, while the later scene (b) feels more open, greener, and calm.

(a) (b)

**Figure 1.1** Street-level images taken from nearly the same location in Delft. The railway was moved underground as part of a major urban renewal project. (a) Before the intervention (circa 2008), and (b) after the intervention (circa 2022). Both images sourced from Google Street View [9].

This example highlights the importance of accounting not only for the physical structure of the built environment, but also for the perceptual and emotional responses it generates. To analyze these dimensions systematically, this thesis will introduce a

conceptual distinction between *components*, referring to the observable physical elements in the image, and *conditions*, referring to the affective or experiential qualities those elements convey. This distinction underpins a typology of image-encoded information, which clarifies the kinds of data that images can offer and how they can be used to better understand the relationship between urban form and experience.

### 1.2.2 Images as data modality

From a computational perspective, images are structured arrays of pixel values, typically represented as multidimensional matrices. An image is composed of a grid of pixels (i.e., numerical values), where each pixel encodes light intensity and color information. In standard RGB image format, this information is distributed across three color channels: red, green, and blue. Figure 1.2 shows an image and its decomposition into a three-dimensional matrix with dimensions corresponding to height, width, and color depth (e.g., three for RGB images). For example, if the dimensions (i.e., $H \times V$) of the image in Figure 1.2 are $640 \times 480$, then the image contains over 300,000 individual pixel locations, each with three color values, one for each channel.

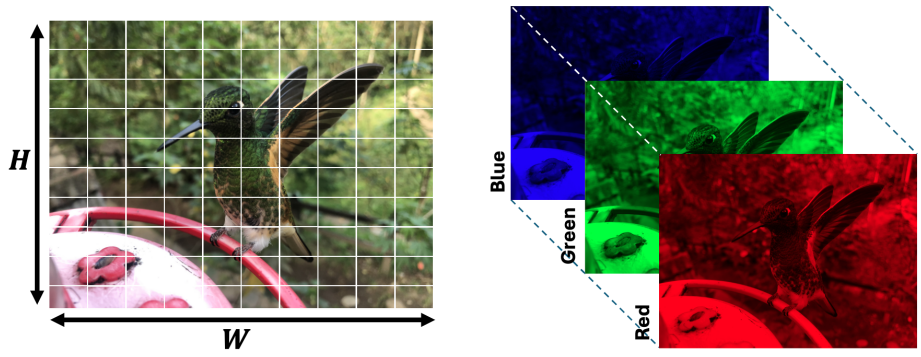



**Figure 1.2** Image decomposition into RGB channels. Each channel maintains the original dimensions of the original image.

These low-level matrix numerical representations form the raw input for computer vision algorithms. However, the semantic value of images relates to what they depict and how they are perceived, is not explicitly encoded in these numbers. Extracting structured and meaningful information from raw pixels requires sophisticated computer vision techniques, which are introduced in the next sub-section.

### 1.2.3 Computer vision

Computer vision is a subfield of artificial intelligence that enables machines to interpret and understand visual information from the world (i.e., images and videos), simulating certain cognitive functions of human vision. CV models convert raw pixel data into

structured, meaningful representations that can inform decision-making or prediction tasks [6]. Even though the field has existed for decades, recent advances in deep learning, particularly the emergence of Convolutional Neural Networks (CNNs) [24], have significantly expanded the capabilities and precision of computer vision systems [25]. Deep learning CV models leverage large annotated datasets to extract multi-level, hierarchical features and concepts from images, which range from low-level patterns to high-level concepts. This hierarchical feature learning forms the basis for most computer vision tasks, which apply these representations to solve specific problems such as classification, detection, or segmentation.

**Models and tasks**

Modern computer vision models typically follow a layered architecture in which visual information is incrementally transformed from raw pixel data into higher-level abstractions. These models are generally divided into two main components: a feature extraction backbone and a task-specific head (see Figure 1.3). The feature extractor, often implemented through deep convolutional neural networks (CNNs) [24] or more recently Vision Transformers (ViTs) [2], processes the input image through a series of convolutional, pooling, and activation layers to build progressively richer internal representations of the visual content. These layers capture low-level patterns such as edges and textures in the early stages, and more abstract features like object shapes or scene structures in the deeper layers [26]. Once this multi-level feature representation is obtained (feature map in Figure 1.3), task-specific layers are attached to guide the model toward particular goals, where the most well-known and popular are classification, detection, or segmentation. These head layers differ depending on the nature of the output: classification heads [16, 25] produce probability vectors, detection heads [12, 27] predict bounding boxes and class scores, and segmentation heads [13] generate pixel-level masks. Figure 1.3 illustrates this general pipeline, along with the outputs produced by each of these three foundational computer vision tasks.

**Image embeddings**

In addition to producing task-specific outputs such as labels or bounding boxes, computer vision models can also generate *image embeddings*. An embedding is a high-dimensional vector that captures the semantic content of data—in this case, an image. These vectors act as compact, abstract representations that preserve relationships between data points in a continuous hyperspace. The key idea is that semantically similar objects are mapped to nearby points in this hyperspace, while dissimilar ones are placed farther apart. This allows models to reason about similarities and relationships in a geometrically meaningful way. This concept was first popularized in natural language processing (NLP) through word embeddings, where models such as Word2Vec [3] mapped words into vector spaces that reflected semantic relationships. In computer vision, the same idea applies to images. Once learned, embeddings enable a wide range of downstream tasks such as similarity search, clustering, retrieval, and classification—often without retraining the original model.

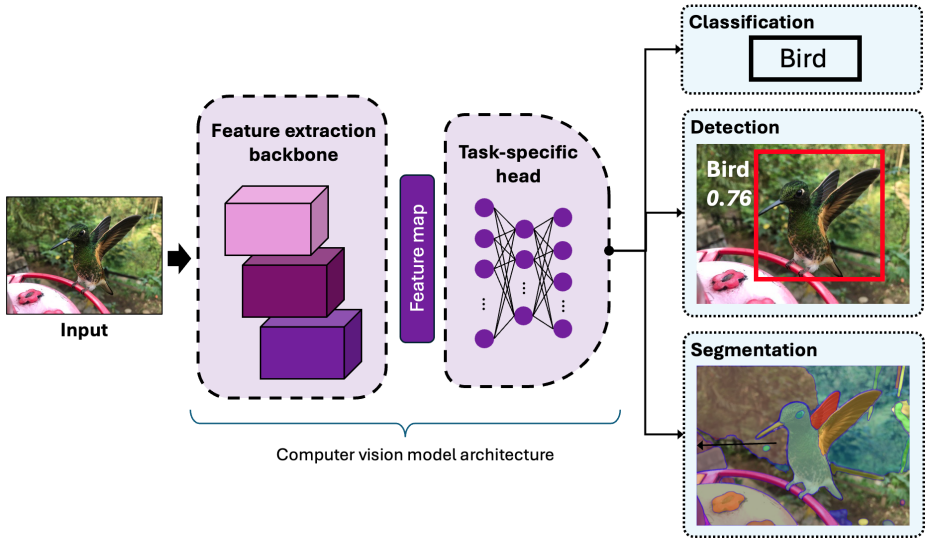



**Figure 1.3** Computer vision general model architecture. The outputs of the three popular tasks are shown: Classification, Object detection, and Segmentation.

To illustrate the embedding concept, consider the example shown in Figure 1.4. Imagine a diverse set of animals, each with distinct attributes such as habitat, size, physical features, and skin type. Suppose we want to organize the animals from the box (a) of the figure using just two semantic dimensions: *size* and *lethality to humans*. By projecting each animal into this 2D space, we obtain a layout like that shown in plot (b). In this embedding, smaller animals are positioned toward the bottom, and less dangerous ones appear to the left. As a result, animals that are semantically similar are located close to one another, while more dissimilar animals are farther apart.

Deep learning embedding models operate on the same principle, but at a much larger scale and with higher dimensionality. Instead of two intuitive axes, the embeddings produced by neural networks typically use hundreds or even thousands of dimensions. Moreover, the axes in these spaces are not easily interpretable as they do not correspond to human-defined categories like size or danger, but the relative distances between vectors still capture meaningful semantic relationships. Modern computer vision models can learn such embeddings directly from raw images, effectively organizing visual information. When human input is incorporated, these embeddings can also reflect perceptual or conceptual similarity.

Embeddings can be learned through several approaches, each requiring different types of supervision and learning objectives. One common method involves the use of *autoencoders* [28, 29], which are neural networks trained to reconstruct their input. This means models receive an input and should be able to return the same input as output. By compressing input data into a lower-dimensional latent representation and

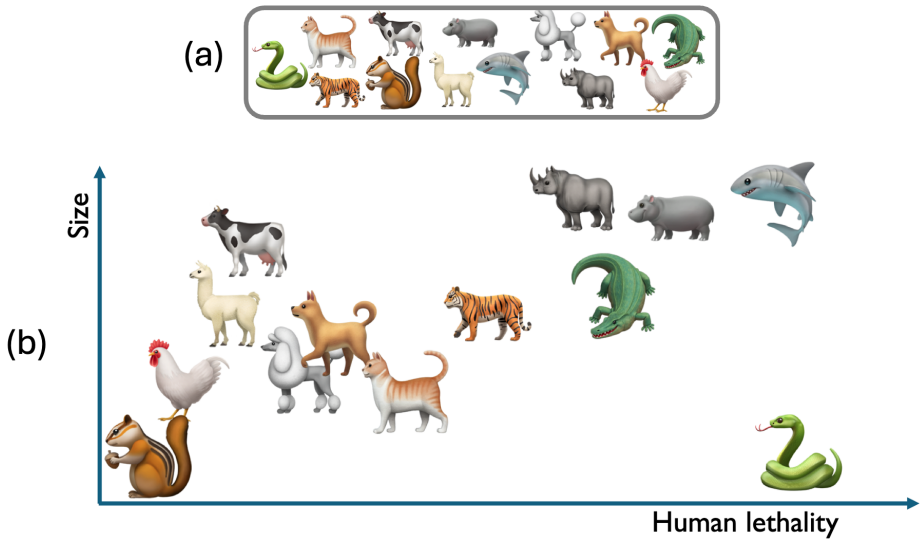



**Figure 1.4** Illustration of an embedding space constructed using semantic dimensions. (a) A set of animals with diverse traits. (b) A 2D embedding space based on the attributes "size" and "human lethality". Similar animals are mapped closer together in the space, while dissimilar ones are farther apart.

then reconstructing it, autoencoders encourage the model to learn compact embeddings that capture the essential structure of the data. A key advantage of this method is that it does not require labeled data, making it suitable for unsupervised learning.

A second approach is *metric learning*, where the model is explicitly trained to position semantically similar items closer together in the embedding space, and dissimilar items farther apart. This is typically implemented using contrastive [30] or triplet loss functions [31]. However, this method requires supervision in the form of similarity information—either labels or pairwise/groupwise similarity judgments—to indicate which instances should be treated as similar or dissimilar.

Finally, embeddings can also be derived from *task-specific models*, such as those trained for classification, object detection, or segmentation. In these models, the intermediate outputs—typically from the last hidden layers—can be repurposed as general-purpose feature representations. For example, in Figure 1.3, the feature map output from the feature extraction backbone corresponds to an embedding of the input image, which is then passed to the task-specific head. Popular architectures like ResNet [25] or Vision Transformers [2] can be truncated just before the final classification layer to extract a fixed-length feature vector for any given image. Once learned, these embeddings can be reused across tasks, datasets, or domains, which makes them a robust and scalable tool for flexible visual analysis.

1

### 1.2.4 Computer vision in urban studies

The use of computer vision techniques in urban research has grown significantly over the past decade. In this context, CV tools have been applied to detect and classify physical elements in cities, or track objects on streets, as well as to analyze human activities, preferences, or perceptions, among other tasks [7]. These applications typically fall into two main categories: (i) the extraction and analysis of physical, observable components from the built environment, and (ii) the approximation of subjective perceptions or experiential conditions. This section briefly reviews both approaches, highlighting their main contributions and the limitations that arise from their conceptual separation.

#### Methods based on components

The first category of studies leverages CV to identify and quantify observable features of the built environment (i.e., components). Core CV tasks in this domain include object detection [32–34], which identifies and locates instances of predefined categories (e.g., pedestrians, trees, or vehicles). Similar approaches have also combined crowd-sourcing and street-level imagery to generate large-scale annotations of urban objects and infrastructure [35]; image segmentation [36, 37], which assigns a label to each pixel in the image, enabling detailed delineation of features; and scene classification [15], which characterizes the overall context or function of a place. Other applications include object tracking [38], which monitors the movement of people or vehicles across time-series frames. A particularly well-established application is the estimation of urban greenery, often operationalized through the Green View Index (GVI), which calculates the proportion of green pixels in street-level images [39, 40]. Similar approaches have been used to study facade conditions and architectural typologies through facade analysis and streetscape mapping [41]. Time-series imagery has also been employed to detect physical transformations in urban neighborhoods, linking visual change to broader socioeconomic dynamics [42]. Together, these studies demonstrate the strength of CV in producing structured, scalable, and objective analyses of urban form, enabling the inference of infrastructure patterns, land use characteristics, and environmental features from visual data.

Beyond classical CV tasks, image embeddings have also played an increasingly important role in urban representation learning. This subfield of urban analytics focuses on generating high-dimensional feature representations of urban spaces from different data modalities, including images. Rather than extracting predefined attributes, embedding-based methods enable more flexible and nuanced encoding of urban scenes, capturing latent patterns that may not be easily described through manual labels. These approaches are often multi-modal, combining imagery with other sources of spatial data such as points of interest (POIs) [43], human mobility trajectories [44], or spatial constraints [45]. The resulting representations can be used for a wide range of downstream tasks, including clustering, classifying urban morphologies, constructing city-level typologies based on visual appearance, or predicting urban-related variables. For instance, learned embeddings can reveal spatial hierarchies or morphological similarities between neighborhoods [46]. This enables meaningful comparative analyses across different

urban contexts. Because these embeddings preserve semantic relationships within the visual data, they are particularly well-suited for tasks that require generalization across cities or regions, even in the absence of detailed annotations.

**Methods based on conditions**

Complementing component-based analyses, a growing body of urban research is using computer vision to explore how environments are perceived and experienced by people. These human-oriented approaches focus on modeling conditions such as beauty, safety, vibrancy, or walkability—qualities that shape how individuals interact with and feel about urban spaces. Unlike physical components, these perceptual attributes are not directly observable; instead, they emerge from people's interpretations of visual cues in context. As a result, perception-oriented modeling depends on human-provided signals—ratings, pairwise comparisons, or choices—linked to specific images.

Several studies have shown that deep learning models—particularly those in computer vision—can approximate such perceptions. Two pioneering works employed image-based surveys in which participants were asked to compare urban scenes along specific perceptual dimensions [22, 47, 48]. These ratings were then used to train predictive models capable of estimating perceptual attributes at scale. Other studies have used convolutional neural networks to predict perceived safety or other perceptions directly from street-level imagery [23]. Other studies have explored crowd-sourcing strategies to collect perceptual judgments from street-level imagery and relate them to urban design characteristics [49]. Additionally, these approaches have also been combined with other methodological domains, such as discrete choice modeling (DCM) [50, 51], to quantify and map perceptions in ways that were previously difficult or impossible using traditional techniques.

Together, these methods signal a shift toward richer, human-centered analyses of the built environment. Perceptions such as comfort, disorder, or enclosure are increasingly treated as measurable and mappable phenomena. CV models—particularly when combined with human input—make it possible to model not only what cities contain, but also how they are seen and felt. This opens new opportunities for research, planning, and design that explicitly incorporate the lived experiences of urban spaces.

**Towards an integrated framework for urban visual analysis**

Although computer vision has become a key tool in urban analytics, integration of physical components and perceptual conditions remains limited. One path has focused on extracting and quantifying physical features of the built environment, while the other one aimed at inferring and modeling how these environments are perceived and experienced by people. These two strands provide valuable but often disconnected insights with limited conceptual or methodological integration [7]. In many cases, the boundaries between them (i.e., components and conditions) are not clearly articulated, partly because the field lacks a shared conceptual framework to distinguish the types of information

that images actually encode [18]. Without this foundation, it remains difficult to design models that treat visual data as both a representation of structure and a stimulus for perception.

Much of the current research in urban CV isolates physical, perceptual, and behavioral layers rather than modeling their interactions holistically [7, 52, 53]. In practice, some studies reduce urban experience to overly simplistic metrics, overlooking the richness and ambiguity of human perception [54]. Similarly, Yu et al. [48] emphasize the absence of scalable frameworks that systematically connect visual components to experiential outcomes. Moreover, while embedding-based methods have become an important tool for encoding physical components from images, their potential for representing perceptual conditions remains largely unexplored. Current perception-oriented research has relied primarily on survey-based models, without leveraging embeddings as a systematic representation of how urban environments are experienced.

Beyond conceptual and methodological gaps, practical barriers also limit integration. There is no widely adopted survey infrastructure for large-scale image-based surveys. Collecting reliable judgments remains costly and time-consuming, with study-specific platforms and sampling strategies that complicate comparability, cultural generalizability, privacy/consent management, and reproducibility. As a result, perception data are scarce and unevenly distributed across cities and contexts. Additionally, there is no standardized typology that clearly separates observable features from inferred qualities. This ambiguity not only limits interpretability but also complicates the development of integrated pipelines capable of identifying and studying both what is visually present and what is perceptually constructed.

Consequently, CV-based urban analytics still tends to offer either a detailed description of what cities contain or a generalized sense of how they feel, but rarely both at the same time. Bridging this gap is essential for models that support design and policy. Integrating components and conditions and developing more comprehensive, human-aware models of urban space lets us attribute perceptions to specific physical features, compare design alternatives within interpretable behavioral models, and transfer insights across cities while respecting local meaning. Future models must go beyond detecting objects or estimating perceptions independently; they should learn how particular features contribute to human interpretations across different contexts, cultures, and use cases.

## 1.3 Research goals

This thesis takes this fragmentation as a point of departure. Although street-level imagery is abundant and computer-vision algorithms are increasingly powerful, urban analytics still lacks an integrated framework that defines what images encode across physical and perceptual layers, prescribes a systematic extraction of data, specifies when and how to include human-in-the-loop in the modeling, and links both layers to interpret cities. This absence leaves planning and urban models blind to important aspects of both the structure of cities and their lived experience. In response, this thesis clearly defines the type of information encoded in images—*components* and *conditions*—develops models geared

towards understanding urban components, and introduces approaches that integrate these two layers within the same framework. In doing so, it seeks to couple **pixels with people and places**.

The main goal of this thesis is to develop, operationalize, and demonstrate computer vision as the methodological foundation for integrated urban analysis—one that connects the physical form of cities with the perceptual, behavioral, and experiential dimensions of urban life.

Concretely, this goal is pursued through four complementary objectives:

(i) **Conceptual clarity on image-encoded information**: clarify and formalize a typology that distinguishes the kinds of information images encode. This distinction clarifies what is directly observable versus what is inferred or interpreted, so that downstream models use the right constructs for the right purposes.

(ii) **Computer-vision methods for urban components**: design computer-vision methods to detect, classify, and represent physical elements of the urban environment across wide spatial coverage, using both task-specific and embedding-based representation-learning approaches.

(iii) **Human-in-the-loop measurement of conditions**: develop a platform that facilitates the creation and deployment of image-based surveys. These surveys are the instruments used to elicit perceptual judgments and preferences linked to specific images or concepts.

(iv) **Integration of components and conditions for urban understanding**: embed both component- and condition-related information within psychological and discrete-choice frameworks to explain and predict perceptions, preferences, and behavior.

Although this thesis operationalizes its goals using current state-of-the-art computer vision architectures (e.g., CNNs and ViTs), the overarching conceptual and methodological framework is intentionally model-agnostic. The theoretical distinction between physical components and perceptual conditions, and the necessity of human-in-the-loop integration, are not tied to the lifespan of current algorithms. As artificial intelligence advances and new visual, multi-modal, or foundational models emerge, they can be seamlessly ingested into the analytical pipelines proposed here. The enduring contribution of this thesis, therefore, lies not in the specific computational models deployed today, but in the structural framework it provides for linking algorithmic visual extraction with human urban experience over time.

1

## 1.4 Research studies

This thesis advances these goals through six interconnected studies that move from conceptual foundations to scalable extraction, and finally to human-centered integration. Figure 1.5 shows a diagram with the six building blocks that complete this thesis grouped by the four research goals presented.

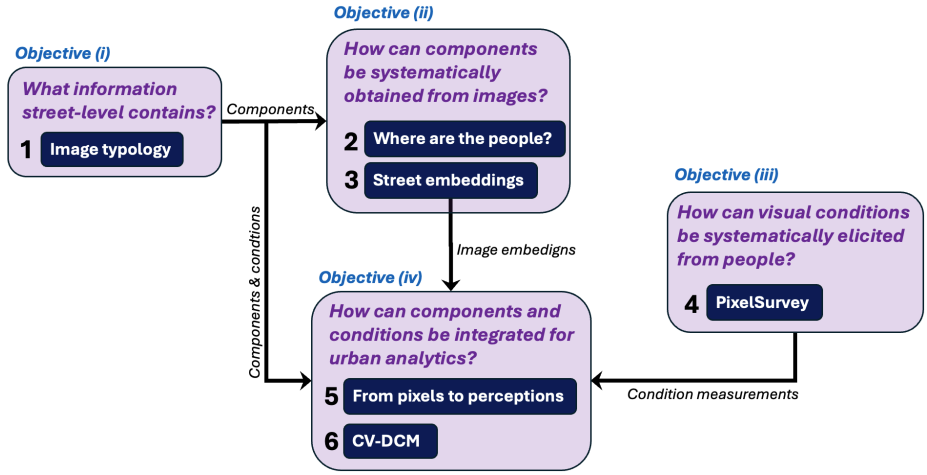



**Figure 1.5** Structure and relationships among the thesis studies.

### 1.4.1 Image typology (Ch. 2)

This research note proposes a concise, landscape-science-grounded typology that distinguishes tangible, observable image content (objects, structures, spatial concepts) from interpretative assessments (e.g., safety, complexity), and provides operational criteria for classification. The typology supplies the conceptual backbone used throughout the thesis to organize what is extracted, inferred, and ultimately modeled.

### 1.4.2 Where are the people? (Ch. 3)

This study develops a large-scale pipeline to detect people in millions of street-level images and relates the resulting spatial variation in "people density" to urban characteristics. It demonstrates how CV can quantify key human components at the metropolitan scale, discusses sampling and bias considerations, and outlines pathways to extend from counts to behaviorally meaningful indicators.

### 1.4.3 Street embeddings (Ch. 4)

In this study, pre-trained visual backbones are used to derive image embeddings for streetscapes. Semi-supervised clustering then yields functional street typologies (e.g., residential, arterial) without handcrafted rules or dense labels. The approach shows that learned visual representations capture meaningful morphological and functional distinctions and are transferable across data-scarce contexts.

### 1.4.4 PixelSurvey (Ch. 5)

To scale the collection of perceptual data, this study introduces *PixelSurvey*, a Python platform designed for image-based questionnaires, stated-choice tasks, and similarity-judgment experiments. It standardizes stimulus management, randomization, and data export, lowering barriers for human-in-the-loop studies that connect images to perceptions and preferences.

### 1.4.5 From pixels to perceptions (Ch. 6)

Building on PixelSurvey, this study elicits human similarity judgments for urban scenes and uses them to align image embeddings with perceptual structure. By supervising representation learning with human similarity (rather than only category labels), the method produces embeddings that better reflect how people group and differentiate places—linking visual components to perceived conditions.

### 1.4.6 Computer vision–enriched discrete choice models (CV–DCMs) (Ch. 7)

Finally, the thesis integrates CV feature maps with Random Utility Maximization to estimate preferences over street-level conditions directly from images, alongside numeric attributes. The model jointly learns a visual feature extractor and preference weights within a logit-consistent utility specification, demonstrates face validity and policy-relevant heterogeneity, and discusses XAI and survey-design implications.

The diagram in Figure 1.5 also makes the dependencies explicit. Objective (i) "Image conceptual clarity" is addressed by the *Image typology* (Study 1), which supplies the conceptual split between components and conditions. The components arrow then feeds Objective (ii) "Scalable extraction of components", where *Where are the people?* (Study 2) and *Street embeddings* (Study 3) deliver scalable measures and representations of physical structure. From Study 3, the image embeddings arrow flows into Objective (iv) "Integration of components and conditions with behavioral models", signaling reuse of those representations downstream. In parallel, Objective (iii) "Human-in-the-loop measurement of conditions" is addressed by *PixelSurvey* (Study 4), whose condition measurements also feed Objective (iv). Finally, Objective (iv) is realized by *From pixels to perceptions* (Study 5), which aligns embeddings with human similarity judgments,

and *CV-DCM* (Study 6), which incorporates component signals and perception-aware features into behavioral choice models. In the following subsection, each study is briefly introduced.

### 1.4.7 Thesis outline

The remainder of this thesis is structured as follows. Chapter 2 (Study 1) sets the conceptual groundwork with the image typology. Chapters 3–4 (Studies 2–3) operationalize component extraction at scale via person detection and street-level embeddings. Chapter 5 (Study 4) presents PixelSurvey, the platform for collecting perception data linked to images. Chapter 6 (Study 5) uses those data to learn perception-aligned visual representations together with components. Chapter 7 (Study 6) integrates components and conditions within computer-vision-enriched discrete-choice models. Chapter 8 concludes by synthesizing contributions, discussing limitations, and outlining directions for future work.

# Chapter 2

# Image Typology


## Abstract

Urban analytics increasingly rely on street-level imagery, yet the field lacks conceptual clarity about what kind of information images encode. This chapter establishes a landscape-science-based typology that differentiates between components—tangible, observable elements such as buildings, vegetation, or people—and conditions—interpretative or perceptual qualities such as safety, beauty, or openness. The typology provides an analytical criterion for classifying visual information by further distinguishing explicit and implicit components, and subjective and objective conditions. This framework clarifies what information images can represent and what can be extracted algorithmically versus what requires human interpretation, creating a conceptual foundation for subsequent computational and perception-based analyses.


This chapter is based on the research note: Understanding environments through imagery — A landscape-science typology of information encoded in images

## 2.1 Introduction

Numerous studies have been conducted to understand how people perceive environments [55–57]. One of the earliest influential contributions to this field was Lynch's [58] three-part model, which elucidated how perception arises from identity (distinct visual objects), structure (recognizable patterns and relationships among objects), and meaning (the emotional values and character of a place). This line of research was further expanded by Gehl [59], who demonstrated how human-scale design fosters social interactions and shapes people's experiences within public spaces. Subsequent research has increasingly integrated psychological, sociological, and technological perspectives to explore environmental influences on perception and behavior [60–62].

More recently, studies into the perception of environments have gained renewed momentum with the proliferation of street-level imagery and AI-driven technologies [63]. The ability to collect and process street-level images has allowed researchers to systematically examine the visual factors that shape human perceptions and behavior across different urban contexts [64, 65]. Moreover, these technological advancements have enabled researchers to scale up studies significantly, allowing for comprehensive analyses across larger and more diverse geographical areas [66].

The growing use of images in landscape science has led to the introduction of new concepts –often drawing upon adjacent research fields– and to changes in how existing terminology is used and operationalized within the field. From the computer vision field, concepts such as feature extraction, object detection, masks, and semantic segmentation have become increasingly prevalent. Similarly, concepts from vision science, including visual acuity, fixations, and saccades, are now widely adopted within landscape studies [67–69]. Concurrently, landscape science's own set of concepts to describe and interpret environments –such as e.g., vista, scene, streetscape, sky view factor, visual complexity, to name a few– has witnessed changes in their use and operationalization. For instance, streetscape, which refers to the arrangement of urban elements such as buildings, trees, and street furniture, is nowadays often represented digitally through street-level imagery and the concept of openness, which measures the proportion of visible sky from a given location, is increasingly assessed using street-level images in combination with segmentation models and operationalized as sky view factor.

However, although the overall meaning of most of the concepts is clear, a well-defined typology for describing the information encoded in images that is relevant to understanding the environments through imagery remains absent. As a result, there is a widespread ambiguity in the use of the concepts, such as image element, object, feature, attribute, indices, metric, and visual dominants – an issue also highlighted in the recent comprehensive review by Liu and Sevtsuk [18]. Oftentimes, studies do not even make a distinction between (visible) elements and those that are perceptual, or they categorize concepts implicitly, without providing a rigorous rationale. This lack of a well-defined typology hampers structured discussions in landscape science studies using images.

To ameliorate this gap, this research note proposes a typology of information encoded in images from a landscape science perspective. Thereby, we hope to provide more clarity to future discussions in landscape studies that make use of imagery.



## 2.2 A landscape science typology for information encoded in images

Our topology, shown in Figure 2.1, asserts that information encoded in images can be understood as consisting of components and conditions. Components are subdivided into explicit and implicit components, while conditions are subdivided into subjective and objective conditions. Below, we formally define these categories, illustrate their use (Section 2.3) and elaborate on boundary cases (Section 2.4).

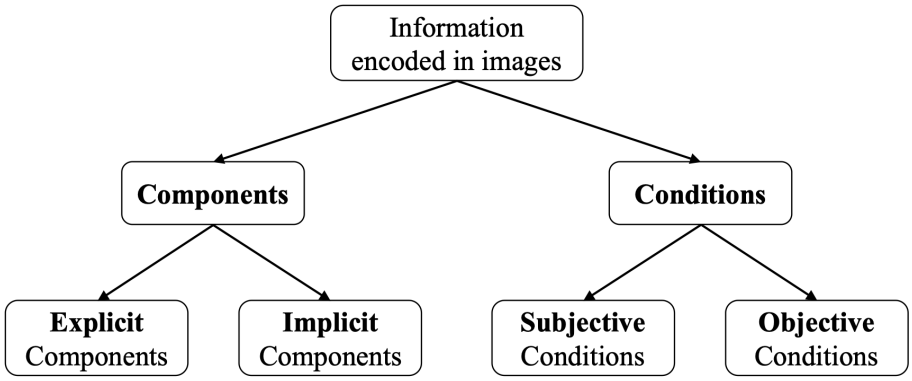



**Figure 2.1** Typology of information encoded in images

### 2.2.1 Components

Components are elements of the scene, encompassing (1) physical objects (e.g., cars, people, trees), (2) structures (recognizable patterns and relationships between objects), or (3) spatial concepts (e.g., sky, horizon, boundaries).

Components can be subdivided into explicit components and implicit components. This subdivision is based on the observation that some components are directly observable while others require contextual reasoning by the onlooker to be identified. For example, a fully visible pedestrian crossing the street is directly observable, whereas a pedestrian largely obscured by a parked car requires reasoning to be recognized as a full person. Similarly, a fully visible tree is directly observable, but if only its shadow and a few branches appear in the image, the onlooker must mentally reconstruct the rest of the tree. Below, we formally define both types of components.

**Explicit components** are components that are largely or fully visible, meeting the following criteria:

- *Direct observability* – An explicit component is visually present in the image and thus can be pinpointed with pixel coordinates without further assumptions about the context. Hence, an explicit component cannot be fully obscured, hidden behind other objects, or exist outside the visible field of view.
- *Pixel-level measurability* – Related to direct observability, an explicit component can be quantified in terms of its colour, shape, location, and spatial arrangement of the elements.
- *Certainty* – The inference of an explicit component's presence does not carry uncertainty. In other words, its presence is non-debatable. Furthermore, the semantic meaning of a component does not rely on the broader context of the image. This means that when a component is isolated from the rest of the image, it keeps its semantic meaning.

**Implicit components** are components that are largely or even fully occluded, meeting the following criteria:

- *Context-based inference* – An implicit component's inference requires reasoning based on contextual cues. These cues may come from spatial relationships, interactions between objects, functional logic, or common knowledge. To put it colloquially, "we can't see it fully, but the broader context strongly suggests its presence".
- *Uncertainty* – An implicit component's inference carries a degree of uncertainty. Some inferences are nearly certain, while others involve a large degree of uncertainty due to factors such as occlusion or limited visual acuity, which may make it difficult to discern finer details.

Components are widely employed in imagery-based landscape science literature, with explicit components more commonly used than implicit ones. In particular, objects such as cars, trees, and vegetation, as well as spatial concepts like the sky, frequently appear in these studies. Their prevalence is largely driven by the capabilities of modern computer vision tools, which are especially effective at detecting such elements.

### 2.2.2 Conditions

**Conditions** are interpretative assessments of the environment. Unlike components, conditions emerge by interpreting the combination of components and require subjective judgment. Their defining criteria are:

- *Interpretative nature* – A condition relies on the onlooker's interpretation of the scene and his/her contextual understanding.

- *Lack of direct pixel encoding* – A condition cannot be traced back to individual pixels, as it emerges by interpreting the combination of components.

Conditions can be further subdivided into subjective and objective conditions. This distinction is based on whether the interpretation involves subjectivity or concerns the inference of a physical or temporal property.



**Subjective conditions** are conditions that concern the subjective qualifications of environments. In the context of landscape science, they often take the form of adjectives applied to the word 'environment'. For example, a person may perceive an environment (depicted in an image) as urban, rural, safe, loud or vibrant. While definitions exist for some of these terms (such as urban/rural and high/low population density), they are inherently subjective and/or context-dependent. For instance, a high population density may be defined as exceeding a certain number of persons per square kilometer, but the threshold chosen is ultimately arbitrary and varies between studies, policy contexts, cultures and time. Consequently, such labels reflect subjective qualifications. Moreover, they tend to vary between individuals and are typically shaped by personal experiences, cultural background, and societal norms.

**Objective conditions** are inferences about physical or temporal properties of the environment that can, in principle, be measured or verified objectively, although not directly from the image itself. Examples include distance, temperature, surface material, noise level, humidity, luminosity, weight, as well as temporal attributes such as time of day, date and season. These physical and temporal properties are not components as they are not directly observable or localized in specific pixels. Instead, they are inferred by interpreting the combination of components and the broader scene context. For instance, while the inference "it's a warm day" is a subjective condition, the inference "the temperature is above 25 °C" qualifies as an objective condition.

In the landscape science literature, subjective conditions have become increasingly prominent in recent years. A review of the literature highlights several conditions that are frequently used, including Street score, Visual complexity, Urban Vibrancy, Visual Walkability Index, Streetscape Diversity Index, and Urban Aesthetics Assessment [**zhang2024**, 18, 63].

### 2.2.3 Illustration of typology

Figures 2.2 and 2.3, respectively, illustrate the proposed typology by presenting examples of explicit and implicit components and subjective and objective conditions using three images.

We start with explicit components (Figure 2.2). In the images, several physical objects are directly observable, such as a tree (Image A), a cyclist (Image B), and a tree trunk (Image C). Despite the objects being partially occluded (e.g., the tree's roots are not visible), the semantic label assigned to each object would remain the same even if the object were isolated from the rest of the image (as shown). Also, a structural element is present in Image C, namely, a line of trees. This structure would still be identifiable if all


**Figure 2.2** Illustration of typology: Components (image source: Google)

**Figure 2.3** Illustration of typology: Conditions (image source: Google)

the trees forming the line were isolated from the image. Finally, several spatial concepts are encoded in the images. Image B features a (blue) sky, while Image C contains a horizon. Continuing with implicit components, we identify a dormer window (Image A), two bicycles (Image B), and truck wheels (Image C). These elements are either occluded or only partially visible, yet their presence is inferred from contextual cues within the image. For instance, in Image B, the presence of two bicycles is inferred based on: (1) the two upper bodies on the cycle path, (2) the presence of another cyclist, and (3) the posture and perspective of the individuals. Another example is found in Image C, where a blue truck on the left has no visible wheels. Nonetheless, we can reasonably infer their presence. But their absence from the image means we cannot be entirely certain.

Moving towards conditions (Figure 2.3), we see that various subjective and objective conditions can be inferred from the images. Images A, B, and C may be perceived as cyclist-friendly, commercial, and rural environments, respectively. These subjective conditions rely heavily on prior experiences of the onlooker (in this case, of the authors) and a mix of visual cues rather than a single component. The environments depicted in these images can be perceived differently by different people. For instance, a person living in a busy area with a similar appearance, as shown in Image A, may consider it busy, while another person may interpret the environment as quiet because little traffic can be seen. Regarding objective conditions, an onlooker might infer from Images A, B, and C that the ambient sound pressure level is below 60 dBA, the ambient air temperature is below 10°C, and the street pole is made of galvanized steel, respectively. While the onlookers' inferences may vary between individuals, each of these conditions can, in principle, be objectively verified.

### 2.2.4 Boundary cases

Finally, it is important to recognize that while the categories defined in the typology are often clear, there are instances where they become less definitive. Therefore, the typology should not be regarded as absolute. For instance, the boundary between explicit and implicit components can be open to debate. In our typology, this distinction hinges on whether the semantic meaning of a component remains intact when it is isolated from the rest of the image. However, this distinction is subtle, and reasonable observers may disagree. Another example concerns color. In our typology, color is treated as a property of an explicit component –after all, the color of a pixel is physically encoded in its RGB values. However, color can also be seen as a subjective condition: an observer interprets an object's color as, for instance, blue or green. The categorization of colors, such as the boundary between blue and green, is inherently subjective and influenced by cultural and linguistic factors [70].

## 2.3 Conclusion

This paper has proposed a structured typology for information encoded in images from a landscape science perspective. The typology distinguishes between implicit and explicit components and subjective and objective conditions. We hope that the proposed typology helps structure future discussions in landscape studies that make use of imagery. In addition, our typology may help guide the development of AI-based methods in landscape science aimed at extracting such information from visual data.

# Chapter 3

# Where are the people?


## Abstract

This chapter develops a large-scale computer vision pipeline to detect, quantify, and analyze urban components. Specifically, it explores the visible presence of people in millions of geo-referenced street-level images across hundreds of cities. A pre-trained object detection model measures human density and examines its spatial association with urban form indicators such as street-network configuration, land-use mix, built density, and other visual urban components. The results reveal robust relationships between compact, mixed-use, pedestrian-oriented environments and higher visible presence of people, while car-oriented or mono-functional areas exhibit lower densities. By visualizing urban activity directly from imagery, the method offers a scalable and reproducible proxy for street vitality—bridging traditional theories of urban liveliness with data-driven evidence. Methodologically, this work demonstrates how computer vision can transform raw imagery into quantitative indicators of human presence, enabling analyses that link physical structure to patterns of urban life. It operationalizes the component layer of the thesis, showing how visual evidence of people can be integrated into models of urban performance and design. The concept component is not formally used in this chapter; however, all the elements extracted from the images are components based on the definition provided in the typology.


This chapter is based on the journal article: Garrido-Valenzuela, F., Cats, O., & van Cranenburgh, S. (2023). *Where are the people? Counting people in millions of street-level images to explore associations between people's urban density and urban characteristics.*Computers, Environment and Urban Systems, 102, 101971.; and the conference paper Garrido-Valenzuela, F., van Cranenburgh, S., & Cats, O. (2022). *Enriching geospatial data with computer vision to identify urban environment determinants of social interactions.* AGILE: GIScience Series, 3, 72. Code is available at the repository: Garrido Valenzuela, Francisco (2025): Data and code underlying the PhD thesis: Pixels · People · Places: Computer Vision and Image Embeddings for Perception-Aware Urban Analytics. 4TU.ResearchData.

## 3.1 Introduction

A thorough understanding of the relations between the people's density in urban spaces and urban space characteristics is essential for urban and mobility planning and, consequently, for policy making [71]. Urban space characteristics concern all city spaces between buildings in the open air [72]. Attaining a better understanding of these relations enables the assessment of the impact of urban developments, identifying patterns of where people tend to be in cities, deciding where to allocate new services, as well as measure the effects of different urban attributes on people's behaviors. Overall, urban planners and policymakers can use these relations to design better cities to attract more people and create livable and inviting urban spaces.



The number of people in urban spaces is (co)determined by many factors, such as time of day, characteristics of places, and weather conditions. Attributes such as urban layout, appearance, number of benches, or traffic were found to influence the number of people visiting a given space [73]. It also has been discovered that people tend to visit places with better walking accessibility [74, 75], greenery neighborhoods [76, 77], places with slow-moving traffic or limited parking [78], and neighborhoods with a shorter distance to the city center, mixed land-uses, and higher densities [79].

However, these studies have been applied mainly with data from small neighborhoods or specific places, which makes it difficult to generalize methods and results. Most of the data in these studies come from questionnaires [80], surveys [75], field observations [81], and paper diary methods [82]. These practices do not allow capturing high-resolution data over large areas because they are often time-consuming, error-prone, labor-intensive, or intrusive. Therefore, larger-scale research and new methods of data collection are needed to better understand the relations between urban space and the number of people.

Fortunately, a large number of studies have developed new techniques for counting people in urban places using different technologies. For example, the location of social media posts has been used to infer the number of people in diverse areas [83–86]. In addition, data from cell phones and Wi-Fi sensors have also been used to measure the movements and number of people in entire cities or regions [87–89]. Recent advances in computer vision also offer promising ways to analyze and collect urban features, human activity data, and people counts from images. Several studies have developed different techniques for estimating the number of people at events and entire cities using images from social media or video recordings [90–92]. These methods can capture a massive amount of high-resolution data over large areas for fine-grained studies.

In this study, we combine the idea of using such new approaches to capture more detailed and spatial-extensive data in order to better understand the relations between the people's density in urban spaces and urban space characteristics. It can be described using a variety of variables, including the road network, traffic volumes, street furniture, land uses, etc. Specifically, we use the widespread availability of geo-tagged images (e.g., from Google street-view or [93]) to create high-resolution datasets of different urban characteristics and human behaviors, and thereby facilitate the analysis of their co-relations. Along these lines, several studies have used street-level imagery for urban

analysis, see the review of [63]. The increasing use of this data source opens the opportunity to also use it to understand how urban crowds relate to urban characteristics. Therefore, the objective of this study is twofold. First, the substantive aim is to deepen the understanding of the relations between the people's density in any urban space and the characteristics of these places. Second, the methodological objective is to develop a computer-vision-based approach for using images as a potential data source to conduct urban studies. The results will provide insights on how different characteristics of the urban space influence the number of people visiting a particular space, which can be used as a basis for urban and mobility planning for the urban areas analyzed. Also, it provides a general method that could be replicated in many cities.



The remaining part of this document is organized as follows. First, the data and their collection are described and explained. Second, the method is presented. Third, we present the case study in the Netherlands. Finally, the results and conclusions are reported.

## 3.2 Data

Two types of data are used in this study. First, Geographic Information System (GIS) data from Open Street Map [94], which includes the location of services and amenities, land-use information, and street networks. Second, street-level imagery from Google Street View (GSV) which corresponds to 360-degree images taken and superimposed on the street network. These two data types are retrieved for each analysis area included in our study.

### 3.2.1 GIS data collection

This study makes use of five different GIS layers: (1) *city boundaries*, which correspond to the geographic boundaries of the city or municipality within which the rest of the data is collected; (2) *street network edges*, corresponding to the streets of the traffic network within the set boundary; (3) *street network nodes*, corresponding to the structural nodes of the street network and intersections; (4) *amenity locations*, which correspond to the services, places, and facilities within the boundaries, such as restaurants, parking lots, or schools; and (5) *land-uses*, indicating the primary land-use for the different sub-regions within the geographical boundaries, such as residential, commercial, or industrial.

All geographic data are collected from OSM using the Python package OSMnx [95]. This package allows obtaining different GIS layers related to the city, such as street networks, location of various stores or services, water areas, and land uses. Municipal boundaries and street networks can be easily obtained using the internal functions of OSMnx. Amenities and land uses can be obtained with OSM tags using a specific key-value. Table 1 summarizes the different layers used and the tag considered when relevant.

**Table 3.1** Summary of GIS layers collected from Open Street Map (OSM).

| Layer | GIS type | OSM tag |
|---|---|---|
| Boundary | Polygon | OSMnx function |
| Edges | Line | OSMnx function |
| Nodes | Point | OSMnx function |
| Amenities | Point | {"amenity": True} |
| Land uses | Polygon | {"landuse": True} |



Amenities and land uses both contain different categories. Firstly, an amenity is defined as a useful and important facility for residents and visitors. Facilities range, for example, from public toilets and public telephones to banks, pharmacies, prisons, and schools [96]. In total, over 100 different amenities can be obtained. To simplify the structure of this data, all amenities are aggregated into nine categories: *food place*, *education*, *transportation*, *financial*, *entertainment*, *public service*, *facility*, *waste management* and *other*. Secondly, a land use describes the main function a land is used for [96]. In total, 37 different land uses can be obtained, where the most important and common ones are: *commercial*, *construction*, *education*, *industrial*, *residential*, *retail* and *institutional*. OSMnx provides access to a wide range of GIS data that can be selected based on the nature of the problem or analysis being conducted. In this research, the most widely used and city-agnostic GIS layer were selected to perform cross-sectional analysis across multiple cities.

### 3.2.2 Street-level images collection

The images we use in this study come from Google Street View (GSV). Images are queried using specific coordinates. Specifically, within each boundaries of a study area, a grid of points is composed, where each point is separated from the others by $d_{grid}$ meters (e.g., 50 meters). Then, for each point, its longitude and latitude information is specified in Google API to extract the surrounding GSV image id. Each GSV image id corresponds to a unique 360-degree panorama view at the street level. Figure 3.1 shows an example of a 360-degree image divided into four individual images based on the rules explained in the next paragraph. In addition, for each image, the date when the picture was taken and its exact coordinates are stored.

Each 360-degree image is divided into four individual images to have more regularity in the angles of view of the images subject to analysis. As shown in Fig. 3.1, a front, back, and two side views of the street are retrieved. To do so, each panorama is associated with the closest street (street network edge) to infer the angle of the street with respect to the horizontal (see angle $\alpha$ in Fig. 3.2). Fig. 3.2 shows an example of image coordinates (red dot) associated with the closest edge to identify the angle $\alpha$. With this angle, four individual URLs are built, corresponding to the four images (Fig. 3.1).

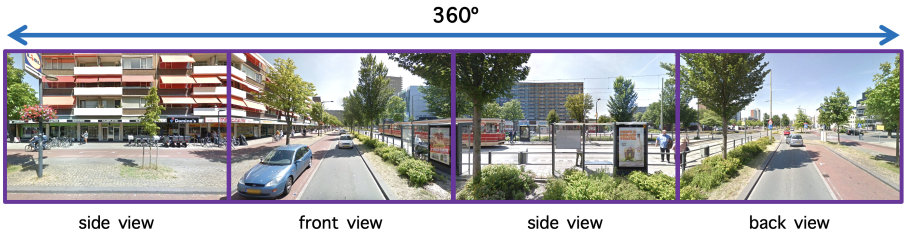



**Figure 3.1** A 360º panorama view retrieved from Google Street View (GSV). Four individual 90º images are shown: front, back, and two side views (left and right view).



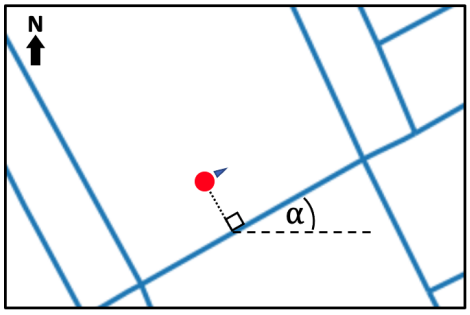



**Figure 3.2** Angle of the street with the horizontal ($\alpha$). The street network is inferred for each GSV image to identify the front view.

Following this data collection process, all GIS data and URLs of the images are stored in databases per study area. Next, the method described in the following section is applied to the data collected.

## 3.3 Method

The method used in this work is divided into two main steps. First, data processing is performed, which includes processing each collected image with an object detection model (ODM) and then aggregating the detected objects and GIS data into spatial units. Second, data analysis is then carried out, which includes the estimation of linear and spatial models to establish the relation between the number of people counted in the images collected within a given spatial unit (*i.e.*, people's density) with the respective urban characteristics in the same unit. Fig. 3.3 shows a diagram of the method which depicts the analysis flow while referring to the data types retrieved in the data collection process.

In the following subsections, steps one and two of the method are detailed.

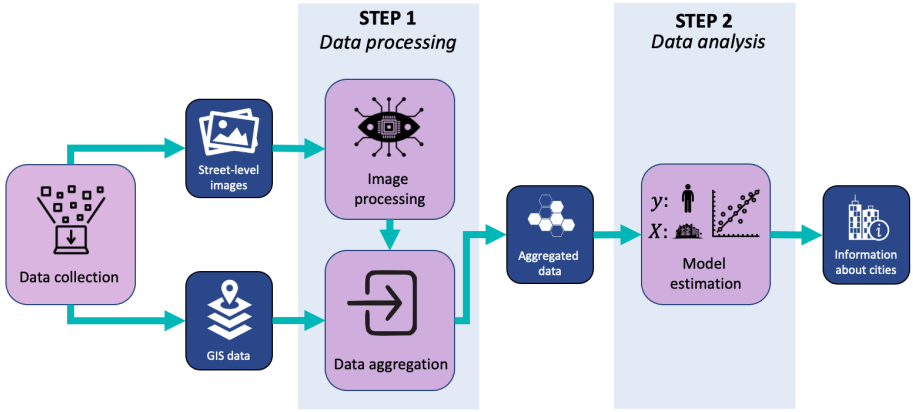



**Figure 3.3** Summary of the method. Purple (large) boxes are sub-steps, and blue (small) boxes are input/output of each sub-step. In data collection, GIS and images are retrieved. Then (STEP 1), images are processed with an object detection model, converted to GIS data, and aggregated in spatial units. Finally (STEP 2), all GIS data is used to estimate various statistical models.

### 3.3.1 Step 1: Data processing

This step has two main objectives. First, it aims to process all collected images to extract information contained in them, which is stored in a GIS data format. Second, it aims to aggregate the GIS data retrieved and the image information in spatial units for subsequent analysis.

#### Image processing

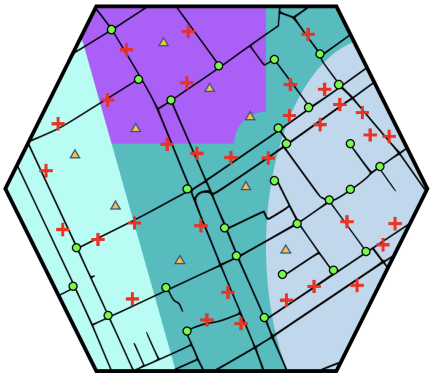
GSV images are analyzed to identify people and urban-related objects. For this purpose, images are processed with an Object Detection Model (ODM) - a machine learning method used to recognize objects in images - to identify people and other urban-related things. Specifically, because this work aims at processing a large number of ideas, the pre-trained SSDMobileNetV3 model [97] is selected. This model is faster compared to other models available at the time of this study for person identification because it was designed to run on smartphones, which requires less computational power. SSDMobileNetV3 is capable of recognizing a large number of objects, but only 13 urban-related objects are selected for this study, namely *person*, *bicycle*, *car*, *motorcycle*, *bus*, *train*, *truck*, *boat*, *traffic light*, *fire hydrant*, *stop sign*, *parking meter*, and *bench*.

Since each image is geolocated, the number of objects identified in each one can be mapped. This means all detections can be stored as GIS data, similar to the data retrieved from OSM. To complement Table 3.1, the detections are stored based on the image's coordinates, registering information on the number of detections per category (*e.g.,* person, bicycle, etc.) in each of the images.

**Data aggregation**

In order to analyze the relations between people counts and urban characteristics, a spatial unit of analysis needs to be defined. For this purpose, regular hexagonal cells with $d_{side}$-meter side (*e.g.*, $d_{side} = 50$ meters) are constructed that tessellate the entire study area. The information collected (see Table 3.1) and processed (from step 1a) within each hexagon is aggregated using different functions. Fig. 3.4 shows an example of a hexagon cell and all possible geographic data contained therein.





| GIS data | Symbol | Attributes |
|---|---|---|
| Edges | | Lenght |
| Nodes | | - |
| Amenities | | Amenity type |
| Land uses | | Land use type |
| Detections | | Detection (person, car, ...) |

**Figure 3.4** An illustration of an individual hexagonal cell. For visual purposes, the length of this hexagon's side is set to 340 meters ($d_{side} = 340$ meters).

*Edges* are aggregated by the sum of the total length of edges within the cell; *nodes* are aggregated by counting the number of nodes per cell; *amenities* are aggregated by counting the total number of places per category per cell (*e.g.*, the total number of educational places, financial places, etc.); *land-uses* are aggregated by the sum of the total area per category per cell (*e.g.*, the total squared-meters of residential area, industrial area, etc.); and *detections* are aggregated by the average number of detections made for the images within the cell per class (*e.g.*, person, cars, etc.). For instance, the aggregated value for the class *person* is obtained by dividing the total number of people detected in all images of a hexagon by the number of processed images within the hexagon.

### 3.3.2 Step 2: Data analysis

Two types of models have been selected to study the correlation between the people's density (*i.e.,* people's counts) in urban places and their characteristics. Both models are applied at the level of hexagonal cells (defined in step 1b). First, a linear regression model is used to infer the impact of each urban characteristic on the number of people present in each cell. Second, a linear model with spatial autocorrelation parameters accounts for the spatial effect of neighboring hexagons on the number of people. In this case, the spatial Durbin model is chosen [98]. In both models, the number of people in the different

spatial units (i.e., hexagons) is the dependent variable ($Y$), and all urban-related variables (e.g., network topology, presence of cars, bicycles, land uses, etc.) are the independent explanatory ones ($X$). In the subsequent subsections, each model is explained.

**Linear regression model**

A classic linear model is used to study the correlation between the number of people identified and the urban-related variables measured within each cell. In Eq. 3.1 $Y$ corresponds to the vector of number of people, where each element is the observed number of people within a particular cell, $X$ is the matrix that contains all explanatory variables (with a constant), $\beta$ represents the vector of parameters for each explanatory variable, and $\epsilon$ is the vector of errors associated.



$$Y = X\beta + \epsilon \tag{3.1}$$

The $k$ explanatory variables ($x_k \in X$) can be divided into four groups: *network* variables which include the number of nodes and total meters of streets, *amenities* and *land-uses* which correspond to the variables previously described, and *detections* which correspond to all variables gathered from image processing. This model aims to identify the associations between all urban variables with the number of people in the urban space, only considering the quantities of each variable. But when spatial variables are studied, near things are more related than distant things [99]. Therefore, another model is used to complement the results.

**Spatial Durbin model**

This model is similar to the linear regression model but takes into account the effect of the neighboring cells' values (spatial correlation) to explain the output (*i.e.*, number of people). Two variables are spatially correlated if they are close to each other and are similar in their attribute values. Specifically, a Durbin model estimation considers the effect of neighbor $Y$ values (number of people in neighboring cells) and the effect of all neighbor $X$ values (*e.g.*, the number of cars in neighboring cells) on the dependent variable ($Y$). In other words, a Durbin model measure spatial auto-correlation (*i.e.*, effect of $Y$ on $Y$) and the spatial correlations (spatial effects of $X$s). In Eq. 3.2, the model specification is shown for a spatial Durbin model in matrix notation.

$$Y = \rho WY + X\beta + WX\gamma + \epsilon \tag{3.2}$$

$Y$ represents the vector of the number of people per cell, $X$ corresponds to the matrix of all explanatory variables per cell, $\beta$, and $\gamma$ are the vectors of parameters for each $x_k \in X$ (with $k$ explanatory variables), linear effect and spatial effect respectively, $\rho$ is the vector of the spatial auto-regressive parameters for $Y$, $\epsilon$ the vector of errors associated, and $W$ is a weighting matrix that measures the effect of neighboring cells. Eq. 3.3 defines each element of the matrix $W$ where the value $w_{ij}$ corresponds to the effect of cell $i$ on cell $j$. In this case, $w_{ij}$ is measured as the inverse of the euclidean distance.

$$w_{ij} = d_{ij}^{-1} \qquad w_{ij} \in W \tag{3.3}$$

In particular, to explain the number of people in a specific cell, this model includes (additional from linear regression) the auto-correlation effect of $Y$ with $WY$ (in $\rho$) and the spatial correlation effect of $X$ with $WX$ (in $\gamma$). It means, $\beta$ parameters take into account the spatial correlations between variables. Therefore, only the $\beta$ parameters are used in the result section (although the spatial parameters ($\rho$ and $\gamma$) provide information on the spatial correlation, unlike a linear model, $\beta$ is already accounting for spatial effects of the $X$s on $Y$.



## 3.4 Case study

The Netherlands is chosen as a case study. Specifically, we use the proposed method to understand the relation between people's density in Dutch urban spaces and the urban characteristics of those places. Also to demonstrate the potential of using street-level images as a data source for urban analytic and behaviors comprehension. The data used in this case study concerns GIS data and street-level images for all municipalities of the Netherlands. As of March 2022, the Netherlands comprises 344 municipalities and has over 17 million inhabitants [100].

Municipalities in the Netherlands vary in terms of surface area and population size. The average surface area is around $97km^2$ and ranges between $[7; 523]$ $km^2$, and the average population per municipality is around 50 thousand inhabitants with a range of $[943; 905k]$ inhabitants.

### 3.4.1 Definition and data collection

Within each municipality, 2022 GIS data and images from different years are collected. The grid used to retrieve GSV images is overlaid using $d_{grid} = 50\ meters$. The smaller $d_{grid}$ is, the more images can be collected, but more computation time is needed for analyzing the images in the posterior steps. With $d_{grid}$ equals to 50 meters we found a good trade-off to have a sufficient amount of data and to allow us to collect all the Netherlands. Next, the data are aggregated in $d_{side} = 50\ meters$ hexagon cells. In this case, $d_{side} = d_{grid} = 50\ meters$ was used to have spatial units with the same level of resolution at which the images were collected. Additionally, a spatial unit with a resolution of 50 meters allows us to capture the local variations in the urban data collected (e.g. land uses) within walkable distances, making it a suitable scale for our research aims. However, the method is able to manage different values of $d_{side}$ and $d_{grid}$ which allows future exploration in multi-scale analysis.

MAUP effects [101] can be generated when the data is aggregated into the 50 meters hexagon cells. MAUP can be separated into two main effects: the zoning effect and the scale effect. Zoning effect refers to the changes in results that occur when the boundaries, shape or position of the areal units are changed. We think this effect has no major implications in our results because we have used a random zonification and

it is composed by a large number of zones (approximately 19 thousand hexagons per municipality in average). On the other hand, scale effect refers to the changes in results that occur when the size of areal unit of analysis is changed. In this case, scale effects can bias the estimation of spatial relationships and patterns. But we think our scale is small enough. Studies such as [102] have shown that greater aggregation zones lead to a decrease in accuracy and precision of the parameters. This means that the lower the level of aggregation (i.e., bigger zones), the greater the loss in efficiency of the parameters. Also, as we mentioned 50 meters can capture local and spatial variations for people which are mostly walking in the urban spaces.



Over 46 million images are collected from 343 municipalities (Baarle-Nassau could not be collected due to issues with border lines between the Netherlands and Belgium). Images are collected from the years between 2008 to 2022. Fig. 3.5 shows a histogram of the number of images collected per municipality on the left and a map displaying its spatial distribution across the Netherlands on the right. It shows that for most municipalities approximately 100 thousands images are collected. The highest number of images are obtained for Amsterdam and Rotterdam (over 1 million images). Based on the rules to collect images (by the traffic network), the number of images per municipality, depends mainly on the surface and the population.

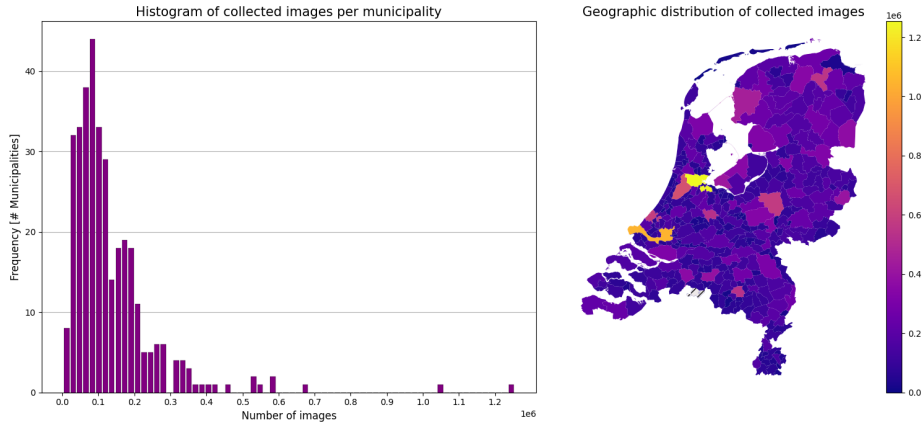



**Figure 3.5** On the left, a histogram of collected images per municipality in the Netherlands. On the right, a map with the spatial distribution of the number of images per municipality

## 3.4.2 Data analysis at different levels of spatial aggregation

Next, we perform three kinds of analysis to explore the associations between people's density in urban places and the attributes of that places at different spatial levels. To do so, the two models presented in the method section are used (linear regression and spatial Durbin model) in three different ways. The analyses are (i) a national aggregated model with all data jointly, (ii) a national analysis using independent municipal models, and (iii)

individual analyses in Rotterdam and Amsterdam. These three analysis considers the 50 meters-side hexagon as minimal spatial unit, but are performed at different scales. The table 3.2 summarizes all analyses and below a detailed description for each is provided.

**Table 3.2** Summary of the three different analyses conducted in this study.

| Analysis | Model | Process |
|---|---|---|
| (i) National aggregated model with all data jointly | Linear regression | Estimation with all hexagons from all municipalities (one model) |
| (ii) National analysis using independent municipal models | Spatial Durbin model | One estimation per municipality (343 models) |
| (iii) Individual analyses in Rotterdam and Amsterdam | Spatial Durbin model | One estimation per city (two models) |



**National aggregated model with all data jointly**

Each 50 meters hexagon of each municipality of the Netherlands is used to estimate a unique Linear Regression Model (see step 2a in the method section). To this end, an ordinary least square estimation is used to obtain the parameters ($\beta$s) of all explanatory variables related to *network*, *amenities*, *land uses* and *detections*.

Since a model is being run at the national level with information collected at the municipal level, it is necessary first to identify which variables can be used to estimate the model. In order to have a robust, cross-sectional result and common variables available across the country, the model is estimated only with the variables present in all municipalities. For example, if there are municipalities that do not have information on military land uses, then this land use will not be included in the national model estimation. Consequently, the following variables have been retained and are included for model estimation: in the *network* variables, the number of nodes and the meters of streets are used; in *amenities*, the number of food places, education places, transportation-related places, financial places, entertainment places, public services, facilities, waste management places, and others are used; in *land uses* only residential, grass area, forest area, and cemetery are used; and for *detections*, bicycle, car, bus, motorcycle, truck, parking meter, and benches are used.

**National analysis using independent municipal models**

Individual models per municipality are estimated. In this case, to take into account the spatial relationships that are inherent to the data, the Spatial Durbin Model is adopted (see step 2b in the method section). By estimating a model per municipality, we can construct a distribution of the different $\beta$s across the country. We, therefore, maintain the same subset of variables as those employed at the national level jointly data version. This allows for the direct comparison of municipal-level models.

**Individual analysis in Rotterdam and Amsterdam**

Finally, the municipalities of Rotterdam and Amsterdam are chosen for a more detailed analysis. The two cities are the most populated cities in the Netherlands and they are geographically close (forty minutes' distance by train). Amsterdam is the capital and it has an urban landscape similar to most Dutch municipalities (the presence of old city centers). Rotterdam, on the other hand, has a different urban and architectural style, following its destruction in World War II. These two cases are used to demonstrate the results of the data processing section and then to compare the estimated model results. This allows for showing the particularities present in the data.



## 3.5 Results

We divide our presentation of the results into two parts. First, the results of the data processing (step 1 of the method) section are exposed. Second, the results of data analysis with the three different approaches (national aggregated model with all data jointly, national analysis using independent municipal models, and individual analysis in Rotterdam and Amsterdam) are presented and discussed (step 2 of the method).

### 3.5.1 Results of image processing step

Street-level images are analyzed to identify people and urban-related objects. Then, all the identified information is aggregated into hexagon-shaped cells. To illustrate how the outcome of data processing looks like, results from Rotterdam and Amsterdam are shown for illustration. Fig. 3.6 shows the spatial variation of the number of people identified within the municipal boundaries of Rotterdam and Amsterdam. As expected, both cities show the highest concentration of people in the respective city center areas. This showcases the possibility of using the information present in images to distinguish between crowded and uncrowded areas.

Fig. 3.7 shows the spatial distribution of the number of private vehicles (one of the independent variables) detected in the municipalities of Rotterdam and Amsterdam. It can be observed that the spatial distribution of private vehicles is more homogeneous than in the case of people (compare Fig. 3.7 to Fig. 3.6). Another pattern that can be observed by visual inspection is that there tend to be fewer cars in places with more people, and vice-versa. This applies for both Rotterdam and Amsterdam.

More generally, our results demonstrate that the outputs from the data processing phase enable the analysis of various urban attributes and their spatial distribution in a study area. The data processing results in this study depend on which models are used to identify objects or situations in the images. In this particular case, we used object detection models to identify a limited number of objects of interest in urban environments. However, this method opens up a wide range of possibilities for gaining new insights by exploring other urban features using perhaps other image-processing tools.

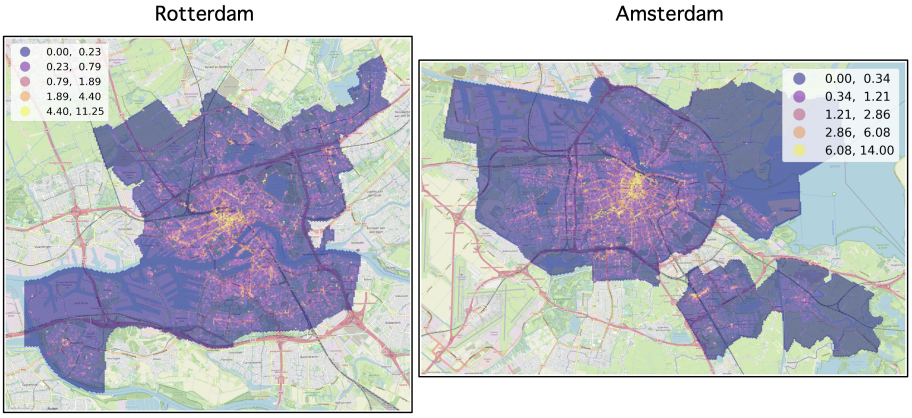



**Figure 3.6** People detection for the municipality of Rotterdam (left) and Amsterdam (right). Values shown correspond to the average number of people observed in images per hexagon and are presented at the 50-meter-hexagon level using a city-specific *natural breaks* color-scale scheme. The numbers in legend show the interval bounds for each color

### 3.5.2 Relation between people's density and urban-related characteristics

One of the objectives of this study is to examine the relations between people's density in urban spaces and the characteristics of those places. This section reports model estimation results for the series of models discussed in the method section (section 3). The results are divided into three parts (as discussed in the case study section): (i) national aggregated model with all data jointly, (ii) national analysis using independent municipal models, and (iii) individual analyses in Rotterdam and Amsterdam.

The analysis of the relation between the number of people (as people's density) and urban-related objects is in each case performed using the 50-meter-sided hexagon as the minimum spatial unit in which the data were aggregated. Municipalities have around 19 thousand hexagons on average, ranging from one thousand to 140 thousand. After removing all hexagons without images, municipalities have about 4 thousand data points (hexagons) on average, ranging between 252 and 18 thousand. An inspection of the deleted hexagons indicates that they mostly correspond to water bodies, agricultural, and natural environment areas.

#### National aggregated model with all data jointly

We estimate a linear regression model using all hexagons with images from the 343 municipalities included. Almost 2 million data points are used to estimate the linear regression model. The model is estimated based on Ordinary Least Squares (OLS). Figure 3.8 shows a bar chart with the values of standardized betas (regression parameters) for each variable. The standardized betas are normalized in standard deviation units,

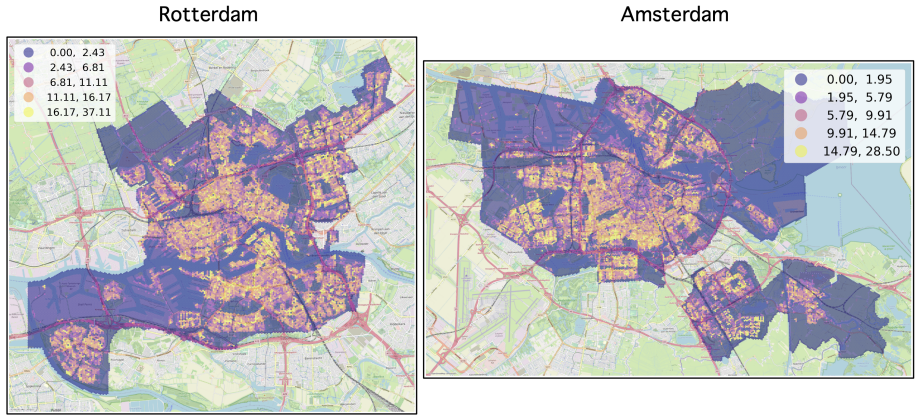



**Figure 3.7** Private vehicles detection for the municipality of Rotterdam (left) and Amsterdam (right). Values shown correspond to the average number of vehicles observed in images per hexagon and are presented at the 50-meter-hexagon level using a city-specific *natural breaks* color-scale scheme. The numbers in legend show the interval bounds for each color

which facilitates the comparison of variables' explanatory power. The goodness of fit index $R^2$ of this model is 25%, which shows that the model is able to explain a substantial portions of the variance.

The national linear regression model results indicate that the number of bicycles detected, the number of food places, motorcycles detected, and the number of nodes (street intersections) have the strongest correlation with the people's density. The strong correlation between people and bicycles/motorcycles can be explained by the mode of transportation used to reach the most crowded areas. As previously mentioned (3.6 and 3.7), where people tend to be, fewer cars are detected. This finding aligns with the results of [78], who finds that less traffic dense areas attract more people. The number of nodes is related to the street network design, indicating that areas of the city with a higher number of intersections positively correlate with the number of people. The number of intersections per hexagon cell could also be related to the block size, indicating that smaller blocks could be associated with more people. This finding is aligned with the work of [103]. She suggests that smaller blocks are more walkable, therefore, more attractive to people. [104] provides empirical evidence that local businesses in dense networks may increase urban vitality. This is in line with our finding that food places and small block areas positively correlate with each other. Lastly, the only two negative relations pertain the two land-use variables, squared meters of forests and grass. These two variables are mostly associated with areas where people do not live, so they are expected to have negative values. Nevertheless, their standardized values have among the lowest predictive power.

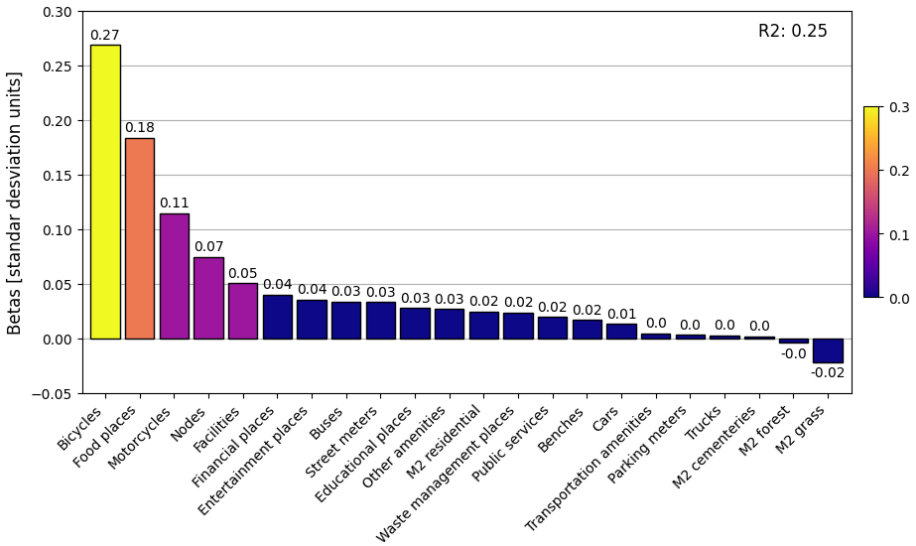



**Figure 3.8** Standardized betas from OLS estimation using all hexagon data jointly. Standardized betas use standard deviation units to facilitate the comparison of the explanatory power of variables

### National analysis using independent municipal models

The model reported in the previous sub-section provides a general indication of the relation between people's density and different urban characteristics. We expect to find different patterns in different cities, potentially allowing us to unravel local relations from the spatial model (SDM). Therefore, it is important to explore how much difference can be found in the explanatory power of each variable across the country. Thus, 343 individual Spatial Durbin models per municipality are estimated, and thereafter, we construct the distribution of the different standardized $\beta$s across all municipalities. These $\beta$ parameters (from SDM) include the spatial effects of the explanatory variables to disentangle the local impact. In addition, we maintain the same subset of variables as those employed at the national level model and the same number of data points per municipality (only hexagons with images in them). This allows for the direct comparison of all municipal-level models between them and the national linear model. Figure 3.9 shows a box plot chart with the values of standardized betas (in standard deviation units) of each municipality for each explanatory variable. The white dot in the middle of each box plot represents the average value of each $\beta$ across the country. The $R^2$ values in these models have an average of $24.5\%$ and range widely between $[6.7\%; 63.8\%]$.

The results of these municipal models point to the same first four variables, which exhibit the most explanatory power as in the national linear model when the box plots are sorted by the median. Figure 3.9 shows that the only explanatory variable which is always positive (for all municipalities) is the number of bicycles detected. Bikes and

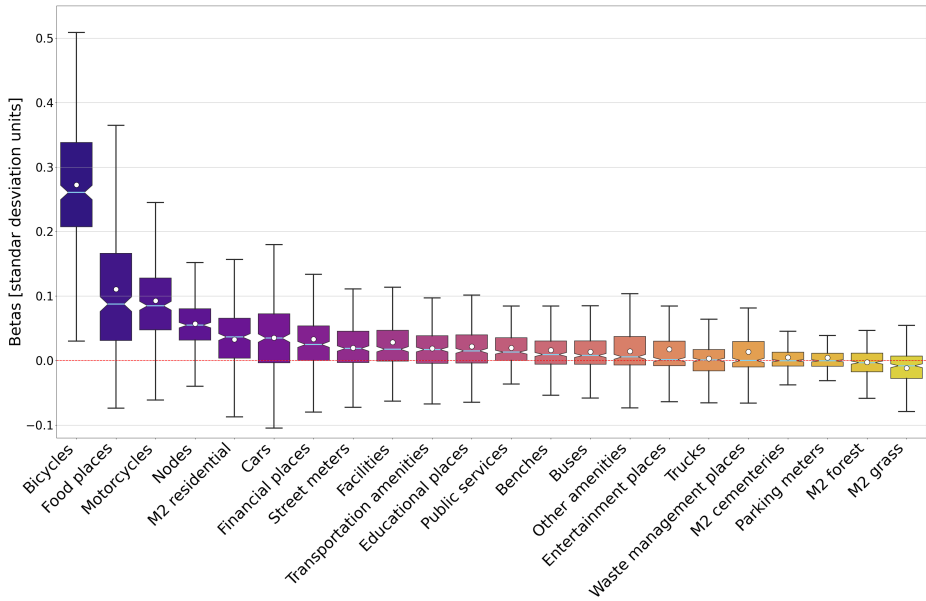



**Figure 3.9** Standardized betas from SDM estimations per municipality, grouped by explanatory variable. The white dot inside each boxplot shows average values.

people tend to be highly spatially correlated across the Netherlands. The number of nodes has again the four highest explanatory power (in terms of median and average) and has a positive relation in most municipalities (only fourteen municipalities show a negative correlation). The results generally indicate a significant degree of diversity in the relations studied across municipalities. This support the idea of study the urban phenomena with a local and small perspective [103], which underscores the importance of studying local effects in order to identify more informative relationships. Additionally, with this results we can determine which relationships are significant when the spatial distributions of the variables is taken into account.

The results presented in Figures 3.8 and 3.9 reveal an at first sight counter-intuitive finding. Specifically, we find "natural places" such as *grass* and *forest* are weakly (negatively) correlated with people's density. An explanation for this small effect can be found in the rural areas. Rural areas contain comparative many "natural places", but tend to be less populated. As a result, few people are counted in these images. This "counter-intuitive" finding highlights an important notion. The space-time accessibility is not accounted for in this study (e.g., [105]). Therefore, the magnitudes of the betas cannot be taken as the isolated effects of the variables (e.g. land-use) on people's density. Rather, the betas represent the strength of the association between the variable and people's density given the spatial and temporal distribution of the variables and images.

The comparison of the national model and the municipal models reveals significant differences in the explanatory power for certain variables, such as cars, square meters of residential land use, and transportation amenities. The national model uses a traditional linear regression and considers data from only one hexagon at a time (one data point), while the municipal models use the Spatial Durbin Model (SDM), which considers spatial effects and data from neighboring hexagons. In addition, by applying one spatial model per municipality, the box plots in Fig. 3.9 can show for each variable the distribution of the explanatory power across the country. For instance, the low explanatory power of cars in the national model is because cars are distributed evenly throughout the country, whereas people tend to be concentrated in specific areas, mainly in city centers. The national model, being a single model for the entire country, may not accurately capture city-specific effects, leading to a small correlation between cars and people. On the other hand, the individual municipal models indicate that some municipalities have a negative correlation between people and cars, while others have a positive correlation, and the median explanatory power for the municipal models is higher compared to the national model. This is because the SDM model can identify relationships between groups of neighboring cells that have similar values, providing a more comprehensive analysis. This highlights the need for local analysis when making policy decisions, as it allows capturing local effects. The following sub-section delves deeper into the analysis for two specific cities.



**Individual analyses in Rotterdam and Amsterdam**

Finally, the municipalities of Rotterdam and Amsterdam are selected for further analysis. In this case, SDMs are estimated to find how the spatial effect of the explanatory variables is related to the number of people in urban places. Figure 3.10 shows the results for Rotterdam and Amsterdam. These results follow the same format as the one used in Figure 3.8, where bars show the standardized beta values. The models for Rotterdam and Amsterdam are estimated with $18,098$ and $17,648$ hexagons (data points), and the $R^2$ values are $33\%$ and $28\%$, respectively.

In terms of differences in explanatory power (standardized betas) of the variables, various differences can be observed between both cities. Rotterdam follows the same pattern observed in the national analyses (previous subsections a and b). Here, the same four variables (food places, bicycles, motorcycles, and number of nodes) appear in the first positions. Compared with Amsterdam, it is a clear difference in the importance of food places. The correlation between the number of people and food places is smaller for Amsterdam. Following this result, Figure 3.10 also shows a higher explanatory power (compared to Rotterdam) for entertainment places and other amenities. Due to the high tourist activity in Amsterdam, the lower explanatory power of food places could be explained because it is shared with these other services (entertainment places and other amenities) in Amsterdam. Related to private vehicles, in both models, the number of cars exercises a negative relationship, corroborating the visual inspection made in the heat maps in Figures 3.6 and 3.7. The negative correlation with cars is higher in Amsterdam than in Rotterdam, which makes sense with the restriction on entering vehicles in the city

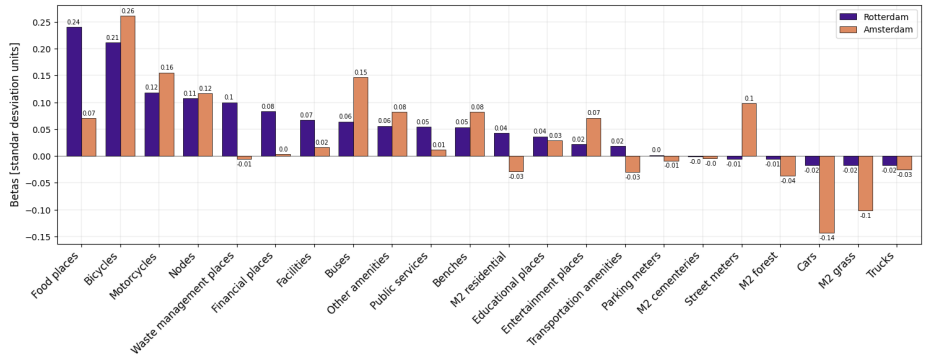



**Figure 3.10** Standardized betas from SDM estimation in Rotterdam (left darker color) and Amsterdam (right lighter color).

center of Amsterdam (where most people are). For the residential land-use variable, a negative relationship is found in Amsterdam, whereas a positive relationship with the number of people is found in Rotterdam.

## 3.6 Conclusions

We have investigated the relations between the people's density in urban spaces and urban characteristics. In addition, we also provide a method for using street-level imagery and GIS data to analyze urban environments.

By processing 46.5 million collected street-level images with an object detection model, we have been able to identify locations where one expects more people's density for selected cases, as well as other objects such as vehicles, bikes, and buses. It seeks to identify in which areas they are frequently observed, giving some indication of urban mobility patterns. Finally, by analyzing the information jointly, *e.g.* location of people and vehicles, it is also possible to perform spatial correlation analysis to identify spatial trends across urban spaces.

Our analysis reveals several interesting substantive results. Firstly, the number of intersections is positively correlated with the people's density. This means that people tend to be in places with a higher number of intersections, which means smaller blocks, suggesting that topological parameters of the network, such as block size, are relevant to this relationship. Secondly, people's density is positively correlated with the number of food places, the number of bicycles, and the number of motorcycles. The same results are found in different of our analyses: a linear national model with the data jointly and national analysis with spatial models by the municipality. Finally, the comparative analysis between Rotterdam and Amsterdam shows an example to discover city-specific patterns such as correlation differences between the presence of cars or food places and the number of people. These kind of substantive results can be used to support

policies that prioritize the design of smaller blocks, as they may increase foot traffic and contribute to a more vibrant and active community in outside spaces. Also, the promotion of the creation of food places in areas with smaller blocks may attract more people. Another policy implication relates to fostering the use of bicycles and motorcycles as a means of transportation, as their presence seems to be positively related to the number of people observed. This resonates with the findings of [78] who concludes that people tend to walk in places where there is less presence of traffic.

This work proposed a new method to use images for studying urban phenomena. Through the use of images, we find it is possible to gather information that are difficult to obtain using more conventional sources of information, such as surveys. In addition, using images is inexpensive and easy to keep up to date. Finally, this method offers a systematic way of obtaining information in the same format for a country or set of cities, allowing systematic comparisons to be made between different places. The urban environment is constantly changing, many of these changes are made by municipalities on a voluntary basis. Our method offers new avenues for measuring how urban projects change the environment and affect people's behavior. Thereby, it can help municipalities, urban planners, and urban researchers to identify which urban characteristics we should pay attention to when planning urban projects.



The main findings of this study align well with those reported by previous research, which use different methodologies. The associations of people's density with food places (an indicator of local/Business places), the number of intersections (as proxy of network density) and facilities (an indicator of urban equipment) are also reported by [104], [106], and [107].

The findings of this study were validated by comparing them with existing research that explored the same relation using different methodologies. The study found that people's density were associated with food places (as local/business places), the number of intersections (as proxy of network density) and facilities (as urban equipment) in urban areas, which is consistent with the findings of other studies conducted by [104], [106], and [107]. The spatial model also reveals the variations of the effects across municipalities, supporting the Jacobs' idea to examine and create policies on a local and small scale for specific regions or cities.

We are aware that this study and the proposed method in particular also has several and important limitations.

Firstly, the method does not consider the variation of the urban characteristics and the concentration of people caused by temporal, seasonal or occasional events. Such variation could bias the relationships uncovered using our approach. Also, the process behind capturing the images by the Google car such as weather conditions during the captures, route chosen by the car or its speed can bias the results generated by this method. To overcome a bit this limitation, we average the detected people in images across years (2008 to 2022) (Note that Google Street View service has a frequency of up to 3 pictures per place per year). As a result, our findings reflect a general trend of people's density in urban spaces and urban elements distribution in cities. We believe that

there is potential for future work exploring the effects of seasonality on people's density and using other computer vision techniques, such as assessing weather conditions, to include its effects in the analysis.

Secondly, the analysis considers objects, such as cars or benches that are detected by our employed computer vision model. But, objects not recognized by this model, such as vegetation or water resources, are not considered in our regression analysis. Moreover, our object detection model does not capture comparatively more abstract urban concepts, such as the condition of the urban infrastructure, parks equipment and vegetation type, to name few. Thirdly, the image database that we use lack samples from certain areas. Google Street-view (GSV) has primarily images of locations accessible by car. As a result, parks, forests, or large open public spaces, have been under sampled. Fourthly, the analysis does not include the socio-demographic characteristics of the people in urban spaces, which may be important for understanding the patterns of people's density. Lastly, it is worth noting that our analysis is not able to make claims about causality, only correlation can be established.



These limitations provide opportunities for future research. First, if temporal and spatial dynamics of people in urban spaces aims to be included, this method could be modified by using other services such as [93], Apple's Look Around, or local dedicated companies such as [108] in the Netherlands or [11] in China. Some of these platforms offer street-level images with a better time resolution to include temporal dynamism. Also some of these platforms cover areas without accessibility by car which are parks, forest or open public spaces. Second, other computer vision techniques can be applied to the images to uncover more urban characteristics. Other detection models such as YOLO [27], PSPnet [109] or Transformers-based detectors [110] or segmentation models [109] can be applied. Even, more sophisticated models to infer perceptions of images (such as beauty or safety) can also be implemented [22]. Third, the temporal component of the images can be included in this kind of study in order to establish causal effects between variables.

Our current work utilizes a data-driven approach. For future work, researchers can replicate our method and apply it in conjunction with other urban theories like central place theory [111] and urban size distributions based on Zipf's law [112]. These theories suggest that urban areas are structured hierarchically around central locations. In addition, further research can be done to investigate the effects of aggregating the data at different spatial scales such as neighborhood, district, or city levels - providing a deeper understanding of how urban environments are organized for different transport modes (*e.g.*, walkable, automotive, or transit-friendly areas) and scales. Moreover, the effects of other characteristics of the urban environment on different human behaviors or activities can also be studied. For instance, the proposed method can complement works about urban environment and physical activity [113, 114], urban mobility [115], covid-related effects [116], characterization of urban spaces based on its functions [117] or other behaviors such as walking dogs [118]. To do so, different behaviors and situations can be

identified in the images and perform similar analyses as presented in this research. In addition, this method could be used to verify policy measures and quantify its effects in the urban environment, such as car-restrictions zones and new bike-friendly spaces.

# Chapter 4

# Street embeddings


## Abstract

This chapter develops an embedding-based approach for representing and classifying urban streetscapes. This is one step further in the urban component layer. Instead of focusing on specific components, this study uses the embeddings as an overall representation and composition of the images. We use a pre-trained convolutional neural network to derive high-dimensional image embeddings from street-level imagery. In this process the models capture (latent) visual features that describe the morphology and components of streets. Subsequently, these embeddings, through unsupervised clustering, yield coherent typologies—such as residential, commercial, or arterial streets—without manual labelling or handcrafted rules. The resulting representations encapsulate the contextual relationships among urban components, providing a scalable alternative to conventional, rule-based classifications. Conceptually, this work extends the thesis' component dimension by encoding the compositional structure of streets into continuous, interpretable forms. Methodologically, it highlights the potential of image embeddings as representation to capture visual, and morphological dimensions of urban space. Again, the concept component is not formally used in this chapter, however, latent dimensions encoded in the embedding are generated based on a pre-trained object detection model. Therefore, they are derived from components.


This chapter is based on an extension of the accepted paper: Garrido-Valenzuela, F., Lange, M., Herrera, J. C., van Cranenburgh, S. & Cats, O. (2025).*An image embedding-based approach for classifying street networks*, Proceedings of the 33rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), Minneapolis, MN, USA. The full version is under review on the journal Environment and Planning B: Urban Analytics and City Science. Code and Data are available at the repository: Garrido Valenzuela, Francisco (2025): Data and code underlying the PhD thesis: Pixels · People · Places: Computer Vision and Image Embeddings for Perception-Aware Urban Analytics. 4TU.ResearchData.

## 4.1 Introduction

Classifying streets is a core component of the design and planning of transport networks. It provides information on the characteristics of the network structure of cities and supports a wide range of practical applications. For instance, such classifications can help define or identify areas that require different driving behaviours for automated vehicles [119] and people [120]; regulate traffic to address safety concerns [121]; and enable targeted transport project allocation [122]. Next to these specific applications, road classification can also shape urban expansion, spatial development [123], and land use patterns [124]. Street hierarchies help guide urban expansion by affecting the placement of commercial areas, residential neighbourhoods, and industrial zones. Additionally, classifications impact accessibility as it may influence and guide business locations and housing demand, contributing to the spatial organization of cities. These factors highlight the importance of a meaningful and adaptable road classification system.



Existing approaches for classifying street networks include manual rule-based systems, statistical algorithms, and data-driven methods, each offering distinct advantages and limitations. Traditional classification approaches often rely on predefined rules related to road function, traffic intensity, or local access levels [125]. These methods typically require individuals exploring the streets and manually collecting information about the relevant attributes. Specifically, many national and urban authorities classify streets mainly based on transport characteristics, with most overlooking the urban context in favor of simpler, more easily interpretable classifications for transport applications [124]. Among other approaches, computational methods have been introduced to extract multiple layers of information for street classification. For example, [126, 127] applied image processing techniques to automate terrain surface recognition for street categorization. Other studies have integrated statistical pre-processing to reduce subjectivity in classification [128], incorporated deep learning techniques to analyze dynamic data such as traffic flow [129], or included accident statistics to refine classification schemes [130]. Most recently, [131] introduced a representation learning-based method to derive clusters from OpenStreetMap (OSM) road infrastructure and point-of-interests (POIs) data. Beyond this, a growing body of urban representation learning studies (e.g., *urban2vec* [43], *hex2vec* [46], *M3G* [45]) has demonstrated the potential of learned embeddings to classify and analyze urban environments. These methods typically rely on large training datasets to build city-specific or domain-specific representations, which showcase the relevance of machine learning methods while also highlighting their data and resource requirements.

Despite recent advancements, most existing street classification methods either rely on predefined street categories or require training models on large, high-quality labeled datasets, both of which limit their adaptability across diverse urban contexts. Many widely used approaches focus on a narrow set of road attributes [124], typically emphasizing functional or traffic-related features while overlooking broader contextual elements such as land use, vegetation, and surrounding infrastructure. Their rule-based nature can make them rigid and less responsive to dynamic urban change, particularly

in rapidly evolving areas. In addition, the need for on-site data collection and expert input can make these methods costly and difficult to scale. More flexible and data-driven techniques have emerged to address some of these issues by leveraging machine learning to model the multidimensional nature of street environments. Yet these techniques frequently require large volumes of labeled and well-distributed spatial training data, as well as intensive training and computational resources, restricting their accessibility and transferability to data-scarce regions. These limitations highlight the need for a scalable, adaptable, and data-efficient classification method that captures the complexity of streets using widely available visual data.

To address these limitations, we propose a data-driven image embedding-based method combined with a clustering technique to classify streets based solely on street-level imagery (SLI). In essence, our approach extracts diverse visual information from images to classify streets in a data-driven manner. Image embeddings are numerical vector representations of images learned by machine learning models. Embeddings capture relevant visual characteristics while preserving relationships between different features. At the same time, SLI provides a widely available, data-rich source of information on road environments. These data capture elements such as terrain type, urban infrastructure, land use, and vegetation. By leveraging pre-trained image embeddings, our approach enables a scalable and flexible classification method that learns directly from visual patterns without requiring predefined rules, additional training, or large labeled datasets. Unlike traditional classification approaches that depend on a small set of road car-focused traffic features, our method captures a broader range of multi-modal transport and urban street characteristics, addressing oversimplifications present in previous frameworks. Furthermore, platforms like Google Street View, Mapillary, and Apple Look Around offer extensive SLI coverage, making this approach applicable even in regions where structured road network datasets are scarce. This automated data-driven tool enhances the capabilities of current frameworks while maintaining a simple implementation. Overall, this approach explores an alternative approach that does not depend on training or manual rules, while remaining adaptable across diverse urban contexts.

The implementation of the proposed method consists of a four-step workflow that classifies road networks using SLI. First, we match geo-tagged street-level images to specific road sections based on the network topology of a designated area. As a result, each section is represented by a collection of images. Second, we extract image embeddings using any pre-trained deep learning model, which encodes visual characteristics into numerical representations. Third, we aggregate the extracted embeddings for each road segment, generating a unique vector representation of the section. Finally, we apply an unsupervised clustering algorithm to group road sections with similar embedding patterns, forming distinct street classes. We apply this method to the city of Delft, the Netherlands, where we collect over 70 thousands street-level images from across the city using the approach described by [132]. These images are used to classify approximately

2 thousands road sections. After the classification, we explore the ability of our method in capturing meaningful distinctions between different road categories such as road type, vegetation, and surrounding infrastructure.

Beyond the main contribution of a street classification method, this study also explores the interpretability of image embeddings for street classification. While image embeddings effectively capture visual characteristics from street-level imagery, they are high-dimensional numeric representations that lack a direct meaning. Traditional classification methods rely on explicit, predefined attributes to categorize streets, whereas embeddings encode complex relationships between features without clearly indicating which aspects contribute most to the classification. To address this, we analyze the clustered road segments by examining different dimensions, including attributes commonly used in existing classification frameworks, such as road surface, type, or speed. This analysis provides insight into the degree to which the embedding-based approach aligns with conventional classification criteria and reveals the types of visual characteristics the model considers when grouping roads, thereby enhancing our understanding of the potential and limitations of using image embeddings for street network classification. The present paper extends a preliminary version [133] where we introduced the pipeline and demonstrated its feasibility in Delft. It incorporates a revision of different street classification methods, interpretability and attribute-based analysis with urban and road characteristics, street hierarchies decomposition, and a broader discussion of implications and limitations for transport and urban studies.



The remainder of this paper is organized as follows. Section 2 provides background information on street classification methods and image embeddings. Section 3 describes the required data and our proposed method for making the street classifications. Section 4 presents the results of applying the method in Delft, including the clustering outcomes, its structure, a visual exploration of the classifications, and an analysis of meaningful attributes of the embeddings within the clusters. Sections 5 and 6 conclude our study by discussing the results, implications of our findings, and potential future research directions.

## 4.2 Background

This section provides background on street classification methods and image embeddings. We first review traditional and data-driven classification approaches, then introduce image embeddings as a tool for capturing the visual characteristics of urban environments.

### 4.2.1 Overview of street classification methods

Street classification plays an important role in transport planning and land use patterns in cities [134]. Classifying streets helps define network hierarchies, guide urban development, and regulate transport policies. Traditionally, street classification has relied on predefined characteristics based on functionality and traffic characteristics. However, recent advances in data collection, machine learning, and computational techniques

have led to more data-driven approaches. In this section, we provide an overview of street classification methods, highlighting both traditional frameworks and data-driven techniques.

**Traditional approaches**

Traditional street classification systems serve as a framework for organizing transport networks based on functionality and traffic flow. These systems have been widely adopted by national and municipal institutions worldwide [124] to guide road design, traffic management, and infrastructure planning. Historically, street classification has relied on hierarchical frameworks centered on car movement, categorizing roads into trunks, arterials, collectors, and local roads, each serving a distinct role within the network [135]. For instance, trunk roads are mainly freeways, while arterials accommodate high-capacity traffic movement, collectors facilitate connections between arterials and local streets, and local roads primarily provide property access.



Over time, some classification frameworks have adopted a more holistic perspective, integrating the "movement" function (related to transport and mobility) with the "place" function (which considers social and cultural activities) [136]. Additional classification models have introduced new dimensions, incorporating street activity patterns [137], as well as vehicle types and speed limits [138]. However, despite these innovations, most countries continue to rely on one-dimensional classification systems [124], preferring simpler, conventional approaches due to the complexity and standardization challenges of multidimensional models. Only a few countries, such as Germany, Mexico, and the UK, have explored alternative classification frameworks that integrate multiple dimensions and consider broader urban aspects [124].

**Data-driven approaches**

Advancements in computational techniques have expanded the possibilities for road classification beyond traditional frameworks. Machine learning and data mining techniques are now being used to automate classification and incorporate multiple dimensions beyond predefined rules. One approach involves mining vehicle sensor data to classify roads based on real-world driving patterns. [139] developed a method to classify UK roads into different categories based on speed, steering behavior, gear position, and suspension movement. Another approach leverages network analysis and geospatial modeling to classify streets. [140] developed a tree-like network classification system that differentiates streets based on topological relationships and connectivity patterns rather than traditional administrative categories. Similarly, [141] introduced a classification model that incorporates land use, urban environment quality, and cycling infrastructure into street classification. These approaches align with the growing interest in multi-modal transport planning and sustainable urban mobility.

Beyond transport-focused classifications, some studies explored alternative classification criteria. For instance, [142] introduced a statistical clustering approach that classifies roads based on traffic noise levels, emphasizing the relevance of environmental factors in urban classification systems. Another emerging direction is the use of street-level

imagery (SLI) and computer vision techniques for classification. [126] proposed a system designed for autonomous driving that classifies roads by analyzing image textures and colors from vehicle-mounted cameras, distinguishing between urban, rural, and highway environments. Similarly, [143] developed an illumination-invariant street classification method that enhances robustness by mitigating the impact of changing lighting conditions in object detection models. These advancements demonstrate how visual data can enhance classification accuracy by capturing rich environmental characteristics. Lastly, [131] introduced *Highway2vec*, a method for generating embeddings of microregions based on their road infrastructure characteristics from OpenStreetMap data. The embeddings facilitate meaningful arithmetic operations and clustering, enabling the development of a high-level typology of urban road networks.

4

### 4.2.2 Image embeddings

Image embeddings are numerical vector representations of images that encode visual characteristics in a high-dimensional space. The concept of embeddings was introduced by [3] in Natural Language Processing (NLP) research as a way to represent words in a continuous vector space while preserving semantic relationships based on the meaning of the words. The idea was later extended to other domains, including computer vision, where deep learning models generate image embeddings by transforming raw image data into structured numerical representations that capture meaningful features such as color patterns, textures, shapes, and visual spatial relationships. The goal of image embeddings is to preserve the similarity between images in this multi-dimensional space, ensuring that visually similar images are positioned close to each other while dissimilar ones are farther apart. Figure 6.1 illustrates the conceptual process and objective of image embeddings.

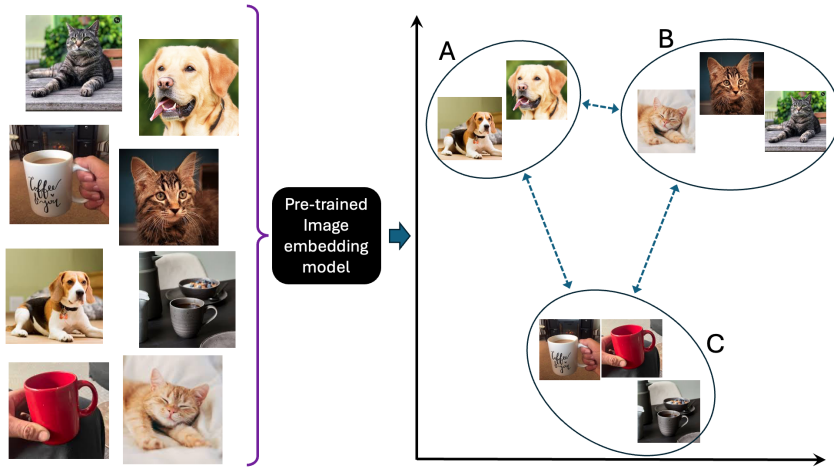



**Figure 4.1** Conceptual process and objective of image embedding models. Images are mapped into a multidimensional space based on their semantic visual content.

Figure 6.1 presents a diverse set of images, including dogs, cats, and cups processed with a pre-trained image embedding model. This model maps them into a high-dimensional space. The result is a structured representation where visually and semantically similar images are positioned closer together. Figure 6.1 projects them in two dimensions for visualization purposes. In the projected 2D space, distinct clusters emerge: dogs (A), cats (B), and cups (C). Notably, the dog and cat clusters are positioned relatively close to each other, reflecting their shared semantic and visual characteristics (e.g., animated, eyes, facial features), whereas cups, being conceptually and visually distinct, form a separate cluster farther away. This illustrates how embeddings capture both fine-grained visual similarities and broader categorical relationships.

Deep learning models, particularly Convolutional Neural Networks (CNNs) and transformer-based architectures, are commonly used to generate image embeddings. CNNs extract hierarchical features from images, progressively capturing low-level patterns (e.g., edges, textures) and high-level semantic information (e.g., objects, scenes, overall composition). Many of these models, such as ResNet [25], were originally designed for classification tasks, identifying objects like cats, dogs, or trees. These models can also be repurposed for embedding generation by removing their final classification layer. Instead of producing class probabilities, these modified models output feature vector representations that encode an image's visual content in a compact numerical form. This adaptation enables classification models and other computer vision models to be applied in broader tasks such as clustering, retrieval, and semantic similarity analysis.



**Applications**

Image embeddings have been widely adopted across multiple domains, including computer vision itself, medical studies, and urban analytics. In computer vision, embeddings facilitate tasks such as image classification, segmentation, and retrieval, improving object recognition and scene understanding. In medical imaging, they support disease detection [144], anomaly identification, and efficient retrieval of relevant medical images [145]. Beyond these fields, urban analytics has leveraged image embeddings for diverse applications. They have been used to classify land use [146], detect and explore urban patterns [43], and analyze neighborhoods [147]. The ability of embeddings to capture a wide range of visual attributes makes them particularly valuable for analyzing the built environment and urban components.

**Relevance to road classifications**

Traditional road classification methods rely on predefined rules and manual categorization based on transport attributes such as road hierarchy, speed limits, and (car) traffic volume. While effective, these methods usually overlook the broader urban context [124], including land use, vegetation, and surrounding infrastructure. Image embeddings offer an alternative approach by encoding the visual appearance of streets directly from images, enabling a data-driven classification method that does not require manually defined categories. Additionally, when pre-trained image embeddings are employed,

they can be directly applied to street classification, even though the models were not originally trained for this context. This is possible because such models have already learned general visual patterns (e.g., textures, shapes, spatial structures) from large and diverse image collections, making their representations transferable to urban imagery without additional training. Moreover, images are currently broadly distributed around the globe, enabling the classification of streets in areas without prior street data. By leveraging image embeddings, road classification can account for both transport-related features (e.g., pavement type, road markings) and contextual elements (e.g., building density, tree coverage, pedestrian infrastructure), making embeddings particularly useful for capturing the complexity of urban environments and providing more adaptable classification schemes.

4

## 4.3 Method

This section describes the input data and outlines our method for classifying the street network based on street-level imagery. Figure 6.3 provides an overview of the pipeline employed in the classification process. To illustrate the method, we include examples from its implementation in Delft, the Netherlands, though the approach is adaptable to any city for which street-level images are available. Delft is a historical city in the western part of the Netherlands, covering approximately $24km^2$. Its size and varied urban landscape, featuring streets from different eras, provide diverse street styles, making it appropriate for testing our road classification approach.

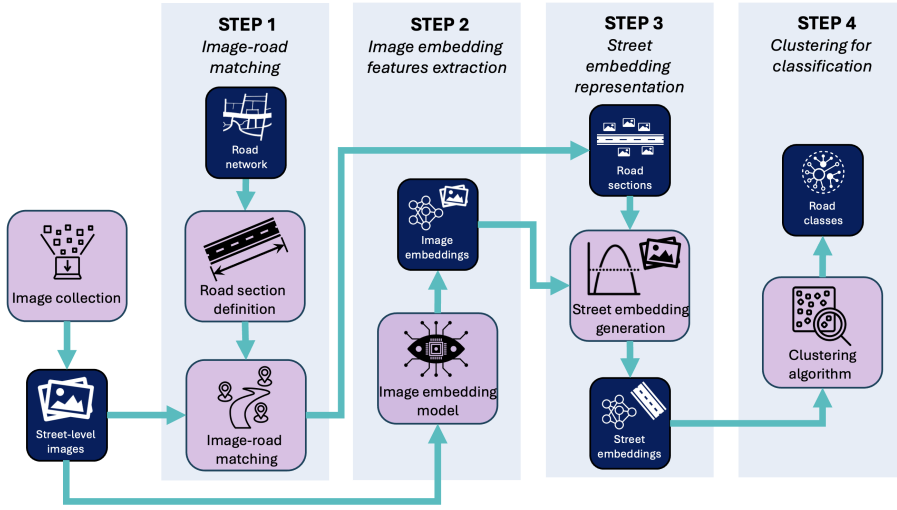



**Figure 4.2** General pipeline of the method.

### 4.3.1 Data

Our classification method utilizes street-level images (SLI) to extract visual street information and a Geographic Information System (GIS) file to represent the road network. SLIs are panoramic photographs captured at ground level, providing detailed visual and structural information about street surroundings. Several providers, including Google Street View (GSV), Mapillary [148], and Apple Look Around [10], offer these images worldwide. For this study, we use imagery from GSV [9], employing the image ID collection method described by [132]. This approach systematically gathers geo-tagged image IDs across a study area, with each ID corresponding to a specific image and its geographical coordinates. We then filter the images to retain only those facing forward and backward along the streets, ensuring that the road and its surroundings are visually captured. In addition to the SLIs, the GIS road network file is available through OpenStreetMap [149] and it provides the coordinates needed to map and identify each street in the study area. For the city of Delft, more than 70 thousands SLIs were collected covering approximately 300 km of road network. Figure 4.3 shows the image ID locations and the network topology for a sub-region within Delft.




**Figure 4.3** Image IDs locations (in red), street network topology (in black) and some road section units (in blue) within a cropped area in Delft, the Netherlands.

### 4.3.2 Step 1: Image-road matching

The first step involves defining and associating the road section units with the SLIs. A road section unit is defined as a spatial line segment that represents either an entire street or a portion of it. These road sections serve as the minimal spatial units for street classification in our method. The definition of a section can vary depending on the application, such as the segment between two intersections, sections based on street names, or fixed-length segments (e.g., 100 meters). Then, the SLIs are spatially associated with the road sections.

In our application for Delft, we define road sections using intersection nodes from the GIS network file. These nodes represent points where streets intersect or experience strong deviations serving as natural boundaries for segmenting the road network. Figure

4.3 highlights six examples of road section units considered in our method, marked with blue circles. The section marked with '*' represents an example of a long segment, which is also split using the nodes from the GIS file. This process yields a total of $3,429$ road sections averaging around $130$ meters each. Once the road sections are defined, we create $20$-meter buffers around them to match SLIs from their surroundings. If an image corresponds to multiple sections, it is assigned to the closest one. As a result, $1,914$ road sections are matched with at least one image, and $68,178$ SLIs are used in total. Figure 4.4 illustrates the coverage, with road sections having at least one associated image in green and sections without images in red.

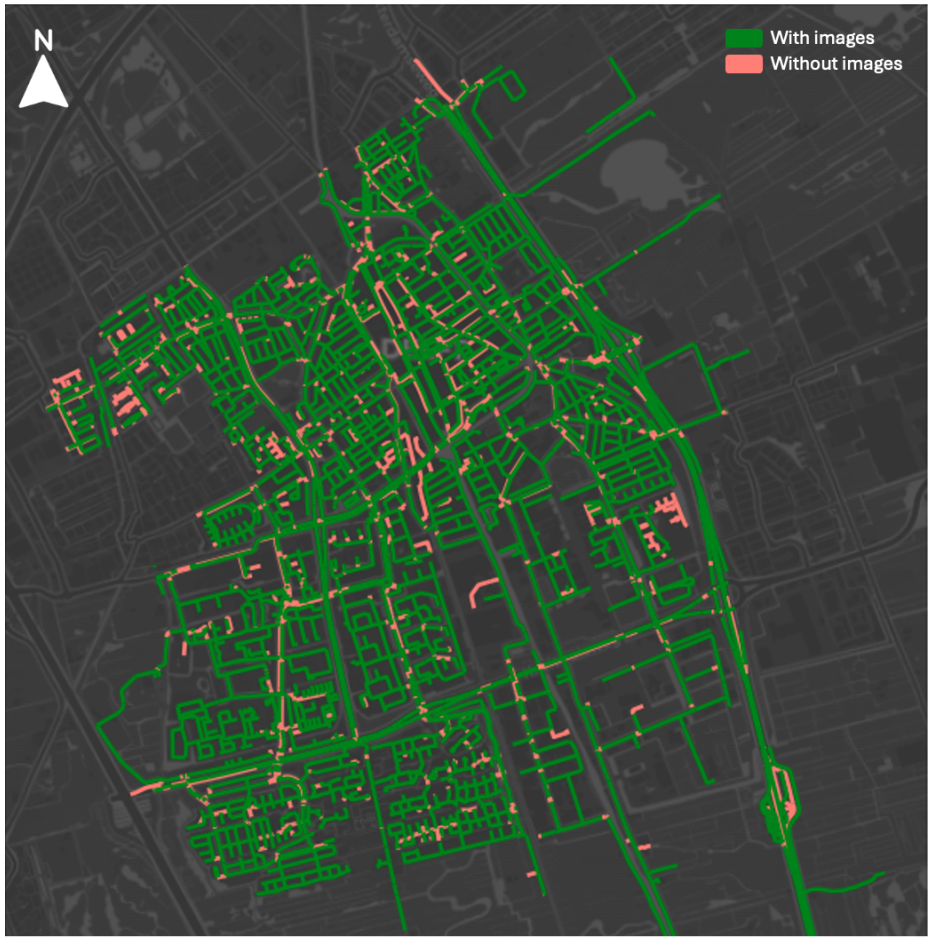



**Figure 4.4** Coverage of images over the road sections. Sections in green are associated with at least one SLI. Sections in red have no images in the 20-meter surroundings.

### 4.3.3 Step 2: Image embedding features extraction

The second step is designed to extract features from the road sections' SLIs. We use a pre-trained image embedding model to achieve this, transforming the street-level images into vector representations. These embeddings capture essential visual characteristics of the images, such as textures, colors, structural patterns, and composition, while reducing the dimensionality of the data. Several pre-trained models can be used for this purpose, including convolutional neural networks (CNNs) and more recent architectures like Vision Transformers (ViTs). Once images are represented as vectors, visually similar images will be positioned closer together in the multidimensional space, as illustrated in Figure 6.1. The distances among images in the multidimensional space facilitate the street clustering based on shared visual characteristics.

We employ ResNet152, a CNN-based model pre-trained on ImageNet [25] with *IMAGENET1K_V2* PyTorch weights [150] for processing the SLI of Delft. We choose this model because of its ability to extract high-quality features from images across diverse contexts and its ease of implementation, making it well-suited for analyzing the visual content of street-level imagery. For each road section, we process the associated images through the embedding model to generate a set of vectors (i.e., one vector per image). ResNet152 generates a $2,048$-dimensional feature vector for each image, preserving various visual attributes relevant for classification.



### 4.3.4 Step 3: Street embedding representation

The third step involves aggregating the image embeddings associated with each road section to create a unified street representation. To achieve this, we compute the mean of the embedding dimensions for all images linked to a given road section. By averaging the vectors, we derive a comprehensive representation that encapsulates the overall visual content of the section, balancing the contribution of each image while reducing noise from individual ones. This approach ensures that the resulting vector reflects the collective visual characteristics of the road section rather than relying on any single image.

Once the street embeddings are generated, we address the curse of dimensionality [151], which can negatively affect clustering performance (in step 4). High-dimensional embeddings can result in sparse data distributions, making the identification of meaningful groups difficult. To mitigate this, we apply Principal Component Analysis (PCA) to reduce dimensionality while preserving at least $80\%$ of the variance. This reduction enhances clustering stability and ensures that the classification results remain interpretable. After PCA, the embeddings are reduced from $2,048$ to 87 dimensions, which balances computational efficiency with information retention.

### 4.3.5 Step 4: Clustering for classification

The final step involves classifying the streets by applying clustering techniques to the aggregated street embedding vectors. The goal is to group road sections based on shared visual patterns captured from street-level imagery. We use the elbow method to determine the appropriate number of clusters, which evaluates the variance explained as a function of the number of clusters and identifies the point of diminishing returns. This approach allows us to balance classification detail with interpretability

For the clustering implementation, we apply hierarchical agglomerative clustering, which is well-suited to the nested structure commonly found in urban road networks. This technique constructs a tree-like hierarchy of clusters, aligning with how streets are often organized, from major roads to local residential streets.



We also test alternative clustering methods to evaluate the robustness of our results, including K-means and Gaussian Mixture Models (GMM). These methods are implemented using the same embedding representations. All three approaches yield consistent classifications, with the core urban typologies remaining stable across methods. Based on this consistency and the interpretability of hierarchical outputs, we select agglomerative clustering for our final analysis.

## 4.4 Results

This section presents the results of applying our embedding-based classification method to the street network of Delft. The findings are organized in two parts. Section 4.1 focuses on the classification outcomes, including the spatial distribution of the clusters, their visual interpretation, and the hierarchical structure among street types. Section 4.2 examines the correspondence between the resulting clusters and the (known) road and urban attributes, providing insights into the types of contextual information encoded by the image embeddings.

### 4.4.1 Embedding-based street classification

This subsection presents the core classification results obtained through our embedding-based method. We begin by presenting the spatial clustering outcomes (4.1.1), which are the main output of our method. Then, we interpret the visual content of each cluster through sampled street-level images (4.1.2). Finally, we explore the hierarchical relationships between street types using a dendrogram structure (4.1.3). These analyses illustrate the model's ability to differentiate meaningful street typologies based on visual patterns only.

#### Clustering results

We apply a hierarchical clustering method over the image embeddings to identify distinct street types in Delft based on their visual appearance. The elbow method determines that six clusters provide the optimal balance between compactness and interpretability. Figure 4.5 shows the resulting classification, with Delft's roads color-coded by cluster.

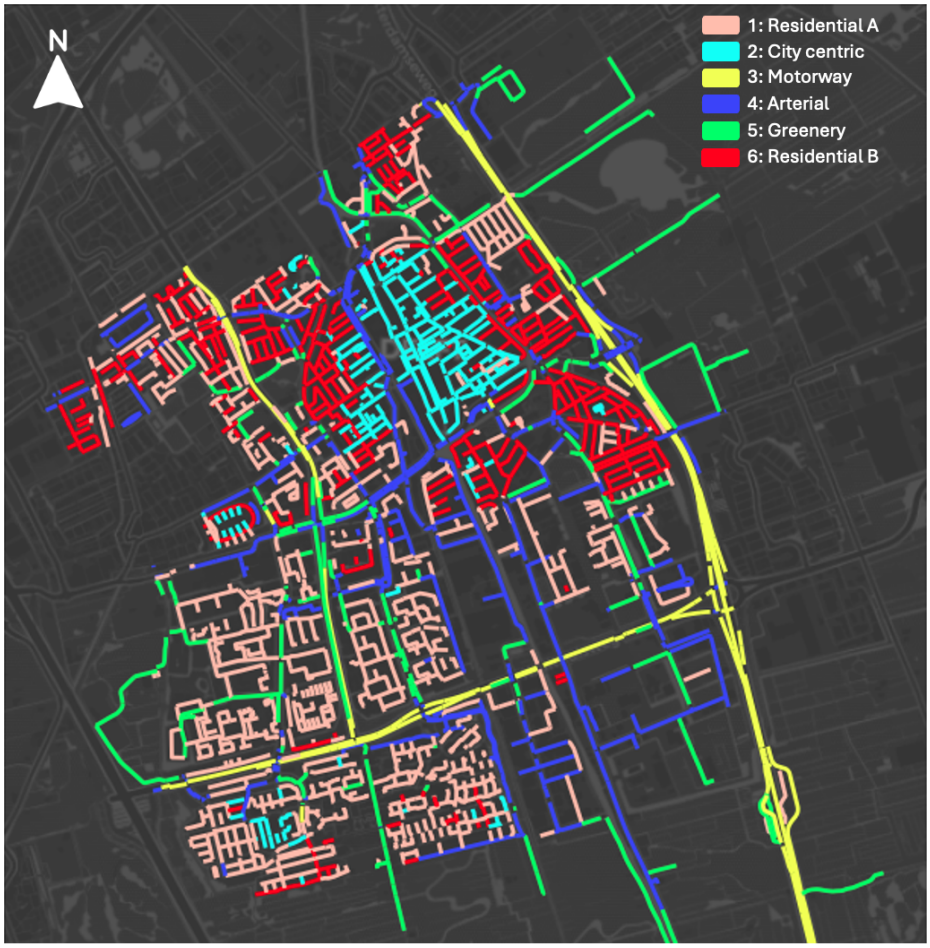



**Figure 4.5** Delft's streets colored by clusters. Clusters names are set based on the visual characteristics of the images.

The model successfully differentiates a variety of urban typologies. The city center, marked in light blue (cluster 2), is characterized by narrow streets, canals, and historic, closely built houses with minimal vehicle presence. Our model accurately identifies the natural boundaries of this area. Notably, the light blue (cluster 2) has two sub-regions separated by the arterial blue road. The smaller one is composed of some streets located west of the train tracks—outside the traditional boundaries of the old town— that are also grouped in this cluster, sharing similar visual features such as street width, facade style, and limited vehicular traffic. On the other hand, a few streets located within the area of the old town (right side of the major light blue region) are assigned to different clusters, likely due to distinctive characteristics such as wider streets or modern architecture.

Additionally, the two major motorways passing through Delft are clearly distinguished in yellow (cluster 3). While the clustering results effectively segment different street types, qualitative validation is necessary to ensure these clusters align with meaningful visual distinctions. To support this, we conduct a qualitative review of randomly sampled images from each cluster.

### Visual interpretation of clustered streets

To better understand the visual differences between clusters and validate their coherence, we conduct a qualitative analysis by sampling random images from each group. This exploratory process complements the cluster definitions by analyzing visual features in the images. Figure 4.6 presents a set of random images from the six clusters.



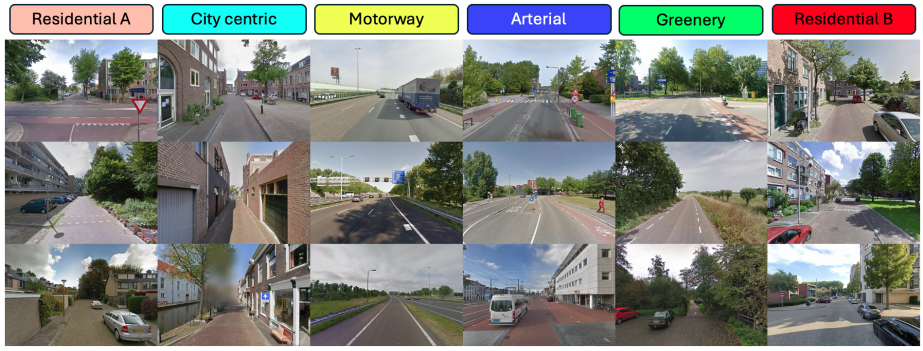



**Figure 4.6** Randomly-sampled images from each cluster.

We labeled the clusters based on the visual characteristics observed in the sampled images and the spatial distribution of the streets in Figure 4.5. Cluster 2 (light blue) is labeled *City Centric* due to its narrow streets and historic buildings. This area closely aligns with the well-known historic center of Delft. Clusters 1 and 6 (pink and red) are labeled *Residential A* and *Residential B*, respectively, characterized by residential houses, open areas, and the presence of trees on the streets. Residential B encircles the city center, maintaining a built environment similar to the historic core. In contrast, Residential A, located in the southern outskirts, exhibits wider streets, newer buildings, and more green spaces, reflecting a more modern residential layout. Cluster 3 (yellow) is identified as *Motorway*, encompassing the main motorways in Delft, and visible from the images showing wide, higher-speed roads. Cluster 4 (blue), labeled *Arterial*, features wider streets with multiple lanes, reflecting their role as major traffic arteries connecting different city areas. Lastly, Cluster 5 (green), labeled *Greenery*, includes roads surrounded by more rural landscapes and vegetation. While similar to Arterial roads in their connectivity function, the visual cues of abundant greenery distinguish this cluster. These labels serve as a preliminary interpretation of the clusters, which can be further refined by incorporating additional data and expert knowledge.

### Hierarchical structure of street types

To investigate the relationships among different street types, we analyze the hierarchical structure of our classification using a dendrogram, as shown in Figure 4.7. This visualization, derived from hierarchical clustering, illustrates how street sections are progressively grouped (from top to bottom in the figure) based on their embedding similarity. In this structure, data points that split higher in the hierarchy indicate broader distinctions, while those that split lower share more similar visual characteristics. This allows us to observe the relationships among street types at different levels of abstraction.

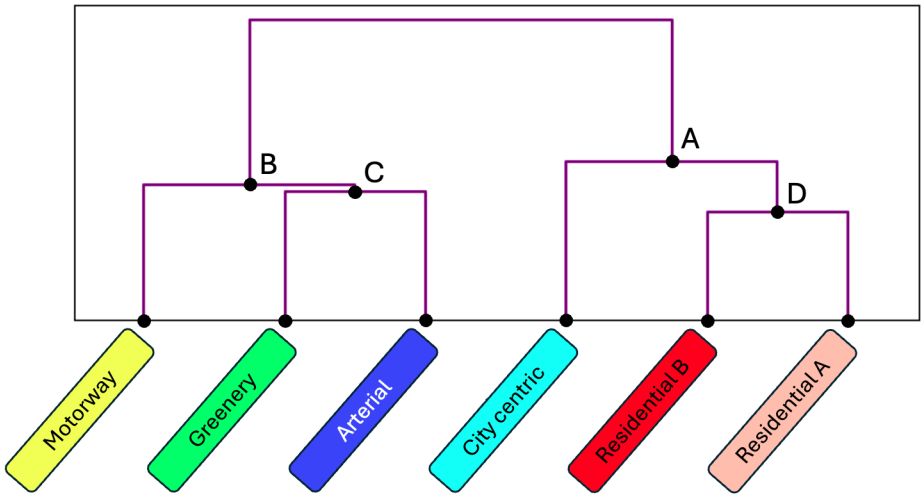



**Figure 4.7** Dendrogram representing the hierarchical clustering structure of street sections.

The dendrogram in Figure 4.7 reveals two primary branches in the classification, where branches A and B are formed. Branch A includes City Centric, Residential A, and Residential B, while Branch B consists of Motorway, Greenery, and Arterial roads. This suggests a clear distinction in their roles within the network, between what can be described as street versus roads. Branch A represents streets with a more residential function, characterized by the presence of buildings, pedestrian activity, and narrow streets. In contrast, Branch B comprises major road transport corridors, where motorways and arterial roads serve as primary routes for vehicle movement, and Greenery roads, despite their different visual characteristics, share a less urbanized environment.

The hierarchical structure further reveals how distinctions emerge within each branch. In Branch A, the first major split occurs between City Centric roads and the two residential classes, highlighting a separation between historic urban cores and general residential areas. This aligns with our previous visual and spatial analysis, confirming that the model effectively captured these distinctions. In Branch B, the clustering structure follows a more direct pattern, with Motorway, Greenery, and Arterial roads separating into their respective categories without additional strong intermediate splits. This

indicates that, although these street types share some overarching similarities, their distinct visual and structural characteristics allow the model to classify them with a high degree of separation, as we can observe in the map (Figure 4.5). This hierarchical structure suggests that embedding-based classification not only distinguishes major street categories but also captures finer distinctions within urban environments. Such insights could inform transportation planning, zoning regulations, and infrastructure development by highlighting the relationships among different street types in a city's spatial organization.

### 4.4.2 Exploring embedding interpretability

In this section, we investigate the extent to which the embedding-based clusters align with established road and urban characteristics. Although our method is based solely on visual data, we examine whether the resulting classifications implicitly reflect attributes such as road type, surface, speed limits, greenery, building age, and land use. The analysis is divided into road-related and urban context attributes, using data from OpenStreetMap and Dutch public datasets.



#### Road-related attributes

We examine three road attributes: road types, surface materials, and speed limits, which have been studied in previous street classification research [126, 127, 138]. Figure 4.8 presents the distribution of these attributes across the six clusters.

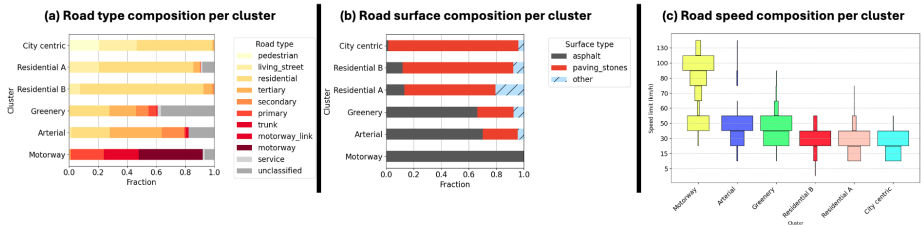



**Figure 4.8** Distribution of road attributes across the clusters. (a) road types, (b) surface materials, and (c) speed limits, across the clusters.

Figure 4.8a shows the distribution of road types across the clusters, sourced from OSM data. The Motorway cluster predominantly consists of trunk roads, motorway links, and motorways, representing the main motorways in Delft. The City Centric cluster primarily consists of pedestrian roads, aligning with its pedestrian-only zones. Residential A and B clusters mainly comprise residential and residential street types. The Arterial cluster shows a blend of primary, secondary, and tertiary roads, indicating its function as a major traffic artery. The Greenery cluster contains more service roads, aligning with its rural nature.

Figure 4.8b displays the distribution of road surface materials across the clusters, focusing on paving stones and asphalt, the dominant surface materials in Delft. City Centric and Residential A/B clusters predominantly feature paving stones, which are characteristic of older streets and lower-speed areas. Conversely, the Arterial and Motorway clusters mainly consist of asphalt roads, typical of major thoroughfares.

Figure 4.8c depicts the road speed limits within the clusters, as reported by OSM. Residential clusters mostly have speed limits of $30km/h$ or less, while Arterial roads often have limits of $50km/h$ or more. The Motorway cluster mainly includes speed limits above $70km/h$. The Greenery cluster (green) shows a mix of speed limits, with a notable presence of roads allowing speeds over $70km/h$, reflecting its rural character.

### Urban related attributes

We analyse four urban attributes: vegetation, building construction year, building height, and land use and amenities composition, as shown in Figure 4.9.

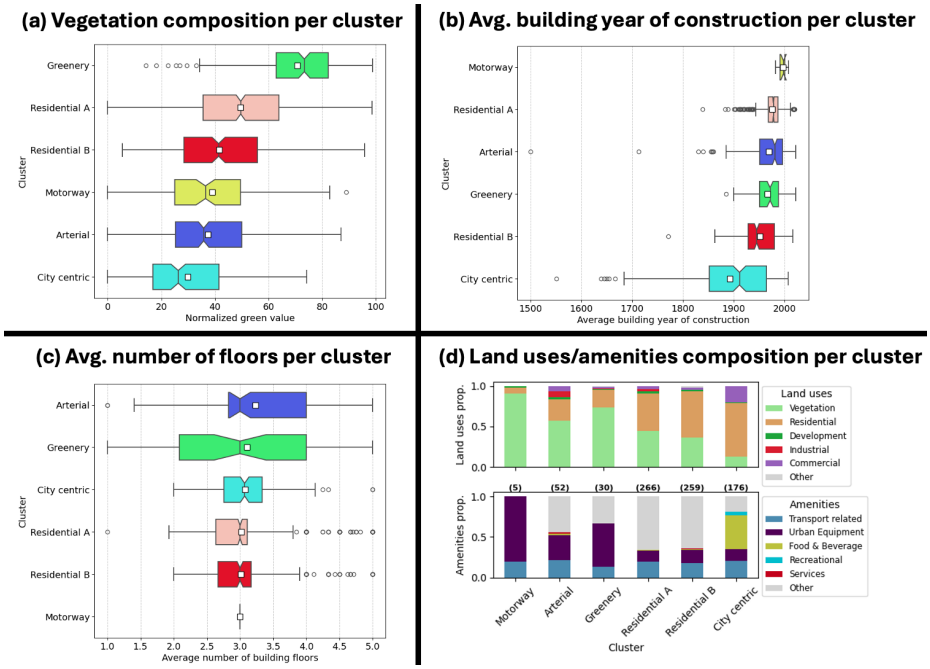



**Figure 4.9** Distribution of urban attributes across the clusters. (a) vegetation composition, (b) average building construction year, (c) average building height (in number of floors), and (c) land uses/amenities composition per cluster.

Figure 4.9a presents the distribution of vegetation across clusters. This data comes from the *Groenkaart* dataset provided by the Dutch government, which assigns a normalized greenery value between 0 and 100 to each location. We compute the average

greenery value within a 5-meter radius of each road section and analyze its distribution across clusters. The Greenery cluster (green) has a significantly higher average greenery value than the other clusters, reflecting the presence of more rural landscapes and vegetation in these areas.

The distribution of building construction years in Figure 4.9b reveals clear historical distinctions. Notably, the City centric cluster (light blue) exhibits the highest variance in construction years and contains the city's oldest buildings, dating back to before 1700. This aligns with the historical nature of Delft's city center, where architectural heritage is well-preserved. Streets in Residential B, which surround the city center, contain older buildings compared to Residential A, suggesting architectural continuity between the city center and its immediate surroundings. Also, this is the major difference between A and B across this analysis. In contrast, the Motorway cluster has the newest buildings with minimal variance, reflecting its development in response to modern infrastructure needs rather than historical urban growth.



Figure 4.9c presents the average number of building floors per cluster. The Residential A/B and City centric clusters show similar distributions, with most buildings between two and four floors, reflecting the predominant mid-rise urban fabric. In contrast, Arterial and Greenery clusters exhibit the greatest variation, with building heights ranging from single-story to five floors. This suggests a mix of residential, commercial, and office spaces, where taller buildings are more likely in areas with major roads or open landscapes.

Figure 4.9d illustrates the distribution of land uses (top) and amenities (bottom) across the clusters. In terms of land use, Residential A/B and City centric clusters contain the highest proportion of residential land use. This reinforces their role as primary living spaces. The City centric cluster also has one of the highest proportions of commercial land use, reflecting its function as Delft's economic and social hub. Meanwhile, industrial land use is concentrated in the Arterial cluster, consistent with the expectation that major roads facilitate industrial and logistical activities. Regarding amenities, the City centric and Residential A/B clusters contain the highest number of amenities (shown on bar tops), confirming their role as activity hubs. The City centric cluster in particular stands out as the most diverse, with a strong concentration of food establishments and recreational services, aligning with its commercial land use. These findings show the multi functionality of the city center, serving both residents and visitors while supporting a vibrant mix of economic and social activities.

In summary, the results demonstrate that our embedding-based method effectively classifies streets into visually coherent and interpretable categories. The clustering approach captures relevant urban typologies, distinguishing between residential, historic city-centric, arterial, and rural environments. Visual exploration confirms the differences of each group, and the dendrogram reveals a logical hierarchical structure. Finally, the attribute analysis shows that the information encoded in the embeddings allows street categorization based on a range of road and urban characteristics, such as traffic function,

greenery, and land use, despite relying solely on visual input. Together, these findings reveal that street-level image embeddings can be used to effectively understand street structures in cities considering different layers of information.

## 4.5 Discussion

We presented a data-driven method for classifying street networks based on geo-tagged street-level images and their embedding representations. By extracting image embeddings and applying unsupervised clustering, our approach classifies streets without relying on predefined categories, extensive road attribute datasets, or high computational infrastructure requirements. This enables scalable, adaptable, and fine-grained street classifications. This responds to calls for more flexible and data-driven street classification systems that move beyond conventional, manually curated hierarchies [124, 125]. The proposed method was applied to the street network of Delft, the Netherlands, resulting in six distinct clusters. Each cluster reflected unique urban typologies and visual characteristics, demonstrating the potential of street-level imagery for detailed road network classification. A key feature of our design is that it reuses pre-trained embeddings and does not train city-specific models, making the approach city-agnostic by construction and therefore transferable in principle beyond Delft. This method has promising applications in urban planning, mobility studies, and automated vehicle navigation [119, 122].



The clustering process revealed clear and meaningful distinctions across the city such as dense historic streets in the city center, different types of residential areas, arterial roads, motorways, and greener, less urbanized roads. By directly incorporating diverse visual information from street-level imagery, our approach addresses key limitations of traditional road classification methods, which often rely on manual input or narrowly defined transport attributes [124, 135]. Visual exploration of random samples within each cluster validated the coherence of the groupings, highlighting that our embedding-based classification captures relevant visual patterns of the built environment. In addition, our analysis of cluster attributes (e.g., road types, speeds, surfaces, land use, greenery) was intended as a second contribution: to explore the type of information generic embeddings capture in practice and the level of alignment of these patterns with conventional criteria.

Furthermore, the identified clusters also reflected differences in road types, speed limits, surface materials, greenery presence, building ages, and land use distributions. This indicates that the visual information embedded in street-level imagery is related to functional, infrastructural, and socio-spatial dimensions of urban streets, extending findings from previous studies on urban form detection using visual data [146, 147]. Additionally, by analyzing the distribution of traditional street and urban attributes across the identified clusters, we explored the interpretability of the image embeddings. This analysis revealed that the visual representations align closely with conventional road classification criteria, offering insights into both the potential and limitations of using embeddings for urban street network classification. Beyond the main classification outcomes, we assessed the robustness of the clustering phase by comparing results

obtained with hierarchical clustering, K-means, and Gaussian Mixture Models. The overall consistency across methods confirmed that our approach is stable and not overly sensitive to the specific clustering technique chosen. Because the method does not involve supervised training, computational cost is dominated by single-pass embedding extraction (feed-forward inference over images) and clustering over segment-level vectors, rather than iterative model fitting. In practice, these steps run on CPUs; GPUs can accelerate inference but are not required for moderate-scale deployments.

We also acknowledge several limitations. The method's reliance on street-level imagery means that image quality, coverage, and representativeness can affect classification outcomes, as suggested by studies relying on visual data under varying conditions [126, 143]. In regions with sparse, outdated, or biased imagery, the results may be less reliable. There is also a practical sampling trade-off at the segment level: too few images risk missing salient characteristics, whereas too many increase cost and noise; establishing principled sampling guidelines, therefore, warrants further definition and evaluation. Additionally, the image embedding process may overlook important road features, such as utilities or temporal changes like construction or seasonal variations, which are not captured in this study. These limitations highlight a common challenge in representation learning approaches, where embeddings capture available information but cannot infer missing context [3]. The clustering process also poses challenges in determining the optimal number of clusters and interpreting the resulting classifications. Future research could explore automated techniques for cluster validation and refinement, as well as methods to improve the interpretability of the clusters. Finally, our empirical evaluation is limited to Delft. Since the method reuses pre-trained embeddings rather than city-specific models, we expect it to transfer to other urban areas; nevertheless, multi-city studies are needed to verify external validity.

Future work could focus on enhancing feature extraction by incorporating more advanced computer vision models, such as Vision Transformers, which have recently demonstrated superior performance in various vision tasks [2]. Comparative evaluations across backbone architectures (e.g., ResNet vs. ViT-based encoders) would clarify performance–cost trade-offs for deployment. Testing the method across cities with different urban forms and development patterns would help assess its generalizability. Practitioner-oriented workflows could enable semi-automatic cluster relabeling by injecting domain knowledge (e.g., thresholds on speed or land use) while preserving the data-driven grouping. Combining street-level imagery with other modalities, such as satellite images or traffic flow data, could further enrich the classification process and provide a more holistic view of road networks [140, 141]. Finally, developing user-friendly tools that translate classification outputs into actionable insights for urban planners and policymakers would help bridge the gap between research and practice [136].

## 4.6 Conclusions

In this study, we proposed a data-driven method for classifying street networks using street-level imagery and image embeddings. Our approach offers a scalable and adaptable alternative that relies only on visual data, addresses traditional rule-based systems' limitations, and reduces data and training-intensive requirements. The classification outcomes reveal meaningful urban typologies—such as residential, city-centric, arterial, and rural streets—derived without predefined labels, high-quality annotated datasets, or deep learning training. Additionally, by analyzing the distribution of conventional urban and road attributes across the clusters, we show that the learned embeddings reflect relevant functional and spatial distinctions. Moreover, this method reuses pre-trained models, involves no supervised training, and leverages widely available imagery, reducing computational burden and simplifying deployment. Although our empirical demonstration focuses on Delft, generic pre-trained embeddings make the approach conceptually transferable across cities. Altogether, this work introduces a lightweight and transferable tool for understanding urban street structures, with potential applications in transport, city planning, and mobility analysis. By combining generic visual embeddings with unsupervised clustering, the approach provides a practical way to scalable street typologies that can be refined with local knowledge when needed.

# Chapter 5

# PixelSurvey


## Abstract

This chapter presents PixelSurvey—an open-source web platform for creating and deploying image-based surveys. The platform enables researchers to create experiments such as similarity judgments, stated choice tasks, or Likert-scale evaluations using large image datasets. In this way, PixelSurvey, acts as a bridge to capture perceptual and affective responses from urban imagery at scale and to incorporate them into computer vision frameworks. PixelSurvey automates key elements of experimental design, including randomization, sampling, and data management, thereby reducing reducing the technical barriers to conducting perception-oriented research. Additionally, it is open source, and provides an easy-to-use, YAML-based configuration system. By offering a modular structure and standardized workflows, the platform promotes reproducibility, comparability, and ethical management of perceptual data. Beyond its technical contribution, PixelSurvey operationalizes the "human-in-the-loop" dimension of the thesis by facilitating the systematic collection of subjective judgments that correspond to the conditions defined in the typology. The platform thus serves as a bridge between computer vision and behavioral science, enabling scalable perception studies that integrate visual, psychological, and urban analytical perspectives.


This chapter is based on an article to be submitted to the journal Behavior Research Methods, Springer Nature titled: PixelSurvey: A modular Python web platform for designing and deploying image-based surveys on preferences and perceptions. PixelSurvey project webpage: https://www.pixelsurvey.io

## 5.1 Introduction

Surveys are among the most widely used methodological tools in behavioral research. They are commonly employed to investigate human preferences, perceptions, and decision-making processes across a broad range of domains. These include transportation [152], urban studies [153], medicine [154], and marketing [155]. Surveys are well-suited for capturing subjective judgments, eliciting stated choices, and assessing attitudes under controlled experimental conditions. The existence of dedicated research fields such as stated choice modeling [156, 157], conjoint analysis [158, 159], factorial survey experiments [160, 161], contingent valuation with dichotomous choice designs [162], and experimental design [163] attests to the relevance that surveys play in the empirical study of behavior. Beyond academic contexts, surveys are extensively used in industry for a variety of purposes, including consumer sentiment analysis, user experience reporting, political polling, and market forecasting. As such, surveys have become an important component in both the scientific investigation and the practical application of behavioral understanding.



A wide variety of software and platforms are currently available for building and managing surveys. Broadly, these can be divided into general-purpose platforms and research-specialized ones. General-purpose platforms (e.g., [164, 165]) are widely used in academia, industry, and government because they are robust, user-friendly, and versatile. They allow users to design questionnaires, deploy them online, and collect responses. These platforms typically support standard question types such as multiple-choice, Likert-scales, rankings, and open-ended items. Research-specialized platforms (e.g., Qualtrics [166], SurveyEngine [167], oTree [168]) go one step further by offering finer control over survey logic, randomization, and experimental design, which makes them particularly suited for behavioral and social science studies. Within this category, psychological and cognitive research often relies on code-based libraries such as jsPsych [169], PsychoPy [170] or lab.js [171]. These frameworks provide maximal flexibility in designing stimulus-based tasks but some of them demand programming expertise. All these platforms provide a rich ecosystem of options that cater to different levels of expertise, study requirements, and methodological traditions across fields such as psychology, sociology, transportation, and marketing.

Alongside this diversity of tools, a growing body of research highlights the central role of visual information in human behavior and decision-making. People naturally rely on visual cues to evaluate environments, products, and experiences [8, 172]. Some recent studies show that image-based designs can also reduce social desirability bias and increase validity in behavioral experiments, particularly in fields such as sociology and political science [173, 174]. As a result, image-based approaches are increasingly used in experimental tasks to assess perceptions [175, 176] or preferences [177, 178] and their potential relevance extends to areas such as travel behavior, environmental economics, and psychology. At the same time, recent advances in computer vision have made it increasingly feasible to analyze, process, and incorporate imagery into behavioral research designs. Deep learning models [2, 25, 179] support classification, embedding,

and retrieval at scale, allowing researchers to incorporate large and diverse visual datasets without extensive manual pre-processing. Therefore, computer vision developments offer new opportunities to incorporate imagery into behavioral research designs, especially in contexts where appearance or visual context are relevant to decision-making or perceptions [178]. As a last point, image-based surveys have been found to be more engaging and enjoyable for participants compared to text-based formats [180, 181]. This not only sustains attention and reduces fatigue but also tends to increase completion rates, thereby lowering the effort required to recruit respondents and achieve target sample sizes.

However, despite the growing relevance of images in surveys, most existing platforms provide limited support for managing experimental designs that rely on large-scale image datasets. On the one hand, some platforms do allow the inclusion of static images as part of survey content. For instance, in [165], [180], and [181], images can be used as alternatives in multiple choice tasks or as stimuli in Likert-scale questions. Yet these tools generally do not allow for a systematic control over image-based attributes, randomization of visual alternatives across respondents, or the combination of images with textual or numerical variables in unified designs [182]. This limitation becomes particularly relevant when images represent explanatory variables and must be sampled, grouped, or presented under specific design constraints. In such cases, researchers are often required to manually process and organize image files, prepare custom randomization scripts, and maintain separate linkages between images, tasks, and responses. On the other hand, research-specialized platforms such as [166] and [167] offer sophisticated experimental logic but are not natively (i.e., without coding) equipped to handle image-based tasks at scale. As a result, deploying image-based survey experiments remains technically challenging, particularly for non-programmers or interdisciplinary teams without dedicated computational support.



Beyond the handling of images, current platforms also remain limited in other relevant aspects. To date, no open-source framework exists that provides a complete environment for designing, deploying, and analyzing survey with experiments. The most similar platform is [165], which is open source, but does not allow one to create experiments guided by experimental designs. This limits transparency, reproducibility, and opportunities for collective improvement. In light of broader concerns about the reproducibility crisis and the growing call for open science practices [183, 184], the absence of open and shareable survey infrastructures is particularly problematic. Lack of transparency not only hampers replication but also increases the risk of questionable research practices or even fraud in survey-based studies [185]. Furthermore, most platforms treat surveys as isolated projects that cannot easily be shared in full—including their structure, logic, and stimuli—for replication or reuse. Finally, survey elements are rarely designed as modular components that can be exchanged or improved collaboratively. This prevents the development of community-driven libraries of reusable modules, such as standardized question types or experimental designs, that could accelerate innovation and foster comparability across studies.

To address this omission in the toolkit of behavioral researchers, we introduce PixelSurvey, an open-source, modular framework for designing and deploying survey experiments. PixelSurvey directly responds to the challenges outlined above. First, it provides a general infrastructure for experimental surveys, where tasks can be flexibly defined through reusable components such as questionnaires and multiple experiment types. Second, it offers native support for image-based experiments, enabling systematic control over large image databases, randomization of visual alternatives, and integration of images with textual or numerical variables in unified designs. Third, it is fully open-source and organized around human-readable YAML "recipes" files. YAML files are simple structured-files that encapsulate the survey structure, logic, and stimuli in a single item. This facilitates sharing, replication, transparency, and adaptation across studies. Finally, its component-based architecture promotes modularity and community-driven extensions, allowing researchers to contribute new question types, experimental paradigms, or designs. Taken together, PixelSurvey lowers the technical barrier to implementing complex survey experiments while extending support for visually rich research designs.



This paper introduces PixelSurvey in detail. We begin with an overview of current platforms and their capabilities and limitations for surveys with experiments. We then present the architecture and workflow of PixelSurvey. This section explains the framework, survey structures, and how surveys are generated from human-readable YAML recipes into fully functional web applications. Additionally, we describe the core components of the framework, including survey sections, activity types, and the supporting database. Specifically, this article introduces three different image-based experiment types which are currently available through PixelSurvey. We conclude with examples of its usage that illustrate how to prepare and build a survey, and discuss opportunities for community contributions.

## 5.2 Existing tools

A variety of software platforms exist for constructing surveys and experimental tasks, each of which with its own distinct strengths and limitations.

### 5.2.1 General-purpose platforms

General-purpose survey platforms, such as LimeSurvey and SurveyMonkey, provide robust infrastructures for questionnaire design, branching logic, basic randomization, and large-scale data collection. These tools are widely adopted in both academic and commercial research due to their reliability and user-friendly interfaces. However, they are primarily designed for conventional surveys with text- or number-based input, and their support for experimental image stimuli is limited. Images can be inserted as static content within questions, and there is no built-in mechanism to systematically manipulate large sets of images or integrate them into complex experimental designs. Additionally, implementing specialized sampling, or advanced randomization schemes is often difficult

or impossible, within these environments. Although LimeSurvey is open source and accepts community contributions, it is not intended for experimental design in behavioral research.

### 5.2.2 Research-specialized platforms

Research-specialized survey platforms have been developed to support more complex experimental designs than conventional questionnaire tools, and are widely used across fields such as psychology, medicine, sociology, political science, and transportation. SurveyEngine [167] is particularly oriented toward stated choice experiments and provides dedicated modules for constructing choice sets and managing experimental designs. However, its support for image stimuli is limited, and it does not allow systematic control over large sets of images within structured experimental frameworks. In sociology and political science, other platforms such as SoSci Survey [186] and Unipark (Questback GmbH) [187] are commonly used to implement vignette and factorial survey experiments, often with complex randomization schemes. But, they face similar limitations in handling large-scale image datasets. Qualtrics [166], in contrast, offers a more flexible environment where researchers can integrate images and extend functionality through custom JavaScript code. This enables the implementation of experimental designs with images. However this comes at the cost of introducing a technical barrier for researchers without programming expertise. Overall, all these platforms are proprietary, which restricts transparency and sharing of complete survey designs. Their commercial nature also means that access can be costly, limiting their adoption for smaller projects or in academic settings where open and reproducible tools are preferred.



For studies that require fine-grained control over stimulus presentation and timing, researchers often turn to specialized experimental software. Tools such as PsychoPy [170] or jsPsych [169] allow precise control of stimulus timing, flexible trial configuration, and the implementation of advanced experimental paradigms. These frameworks are central in psychology and cognitive science, where millisecond accuracy and complex stimulus logic are essential. They typically run as desktop applications or JavaScript libraries and can present stimuli in controlled sequences with rigorous timing. Other open-source frameworks such as oTree [168] and lab.js [171] have become increasingly popular in behavioral economics, psychology, and sociology for building interactive or vignette-based survey experiments online. While all of these tools expand flexibility and support reproducible experimental designs, they are not primarily designed to integrate survey structures with large-scale image stimuli. Additionally, some of them (i.e., PsychoPy and jsPsych),generally requires scripting or programming skills.

### 5.2.3 Statement of need

Researchers who wish to study human preferences and perceptions systematically through experiments (with images) face a clear trade-off. On one hand, general-purpose survey platforms offer accessibility and scalability but lack the flexibility needed for experimental designs and large-scale image management. On the other hand, research-oriented

frameworks deliver experimental rigor but require considerable programming expertise and are not designed for integrating survey structures with image stimuli at scale. As a result, current tools either compromise ease of use or methodological power, leaving researchers without an accessible solution that supports both large-scale image handling and experimental control. PixelSurvey addresses this need by providing an open-source, modular framework that integrates user-friendly survey functionality with fine-grained experimental design. It enables researchers to incorporate large image datasets, systematically manipulate and randomize them, and combine them with textual or numerical variables within unified designs. In doing so, PixelSurvey lowers the technical barrier associated with implementing complex survey experiments, expands the methodological toolkit available for studying human perception and decision-making, and facilitates open science practices.

## 5.3 PixelSurvey



PixelSurvey is an open-source, modular, and community-oriented framework for creating and deploying image-based surveys. Its development is guided by four core principles. First, it is fully open source, ensuring transparency, reproducibility, and enabling community-driven contributions. Second, it adopts a modular architecture, in which survey elements are defined as independent components that simplify future extensions and can be flexibly modified. Third, it is designed to be user-friendly, lowering the technical barrier for researchers who wish to implement visually rich survey experiments without extensive programming expertise. Finally, PixelSurvey is designed to be community-based. We aim to foster collaboration, sharing survey templates, and collective improvement of the platform. Together, these principles make PixelSurvey a unique and versatile framework for collecting human perceptions and preferences. While PixelSurvey's primary motivation was to enable systematic management of image-based stimuli, its flexible architecture now extends far beyond images, making it equally suitable for experiments built entirely on textual, numerical, or other data inputs. In the following subsections, we describe its architecture, workflow, main functionalities, and illustrative use cases.

### 5.3.1 Architecture

The central component of PixelSurvey is **PixelSurveyCore**, which provides the main functionality of the framework. PixelSurveyCore receives a *survey recipe* as input and, based on this configuration, automatically generates a web application. Figure 5.1 illustrates this process, in which a *survey recipe* is used to generate the web service with its own database to collect responses.

The survey recipe consists of two elements: a YAML file (i.e., recipe file) and a source folder (both detailed later). The recipe file specifies the survey structure, experiments, and settings; and the source folder contains all additional files referenced in the YAML such

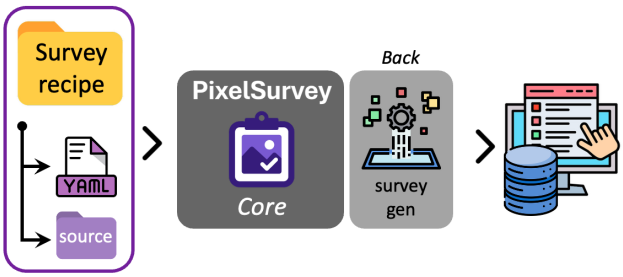



**Figure 5.1** General PixelSurvey pipeline.

as experimental designs or logos. This architecture ensures that surveys are represented in a single YAML file, making them clearly specified, easy to share, and reproducible across different configurations.



### 5.3.2 A survey in PixelSurvey

PixelSurvey adopts a standardized survey structure that is shared across all implementations. This design reduces the complexity of building surveys. Each survey generated with this framework is organized into three sections: On-boarding, Body, and Closure. This structure provides a modular template that improves the readability of survey recipes. In other words, the sections act as an organizational (not functional) framework that facilitates survey design and comprehension, but the actual logic of the experiment is defined at a more granular level (i.e., pages).

Each section is composed of one or more pages, which represent the actual functional units of a survey. These pages correspond to individual web pages encountered by participants and may differ in purpose and content. Although this structure allows for flexibility, it also introduces a certain degree of rigidity to ensure coherence and standardization. Figure 5.2 shows an overview of the sections and their pages, and the following subsections describe the role of each section in the survey.

#### On-boarding section

The on-boarding section always contains three pages: Home, Consent, and Screening. These pages are intended to introduce participants to the study, ensure ethical and legal compliance, and confirm eligibility. The Home page typically presents a welcome message and general information about the survey. The Consent page provides the informed consent statement, outlining participant rights and data protection policies, and requires explicit agreement before proceeding. Finally, the Screening page allows researchers to include questions (e.g., demographics, prior experience) to determine whether participants meet predefined inclusion criteria or to proceed with a representative sample of a target population.

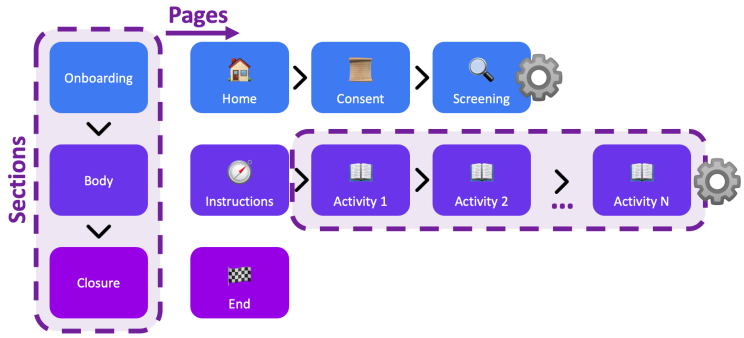



**Figure 5.2** General sections and pages' structure of a survey in PixelSurvey. Blocks with gear icons indicate components that include questions/experimental content and can be customized by users.

While users can freely modify the content of all survey pages, the Screening page has a special role. It allows researchers to include any number of multiple-choice questions that can be used to filter participants and facilitate data collection through demographic quotas or sampling strategies that aim to represent a larger population. This flexibility allows the Screening page supports questions customization and enables targeted participant recruitment.

### Body section

The body section constitutes the central part of the survey. It always begins with an Instructions page, which provides participants with task-specific guidance and clarifies the expected responses. This is followed by one or more activities, which can be either questionnaires or experiments. Figure 5.3 shows the current components for each of those activities.

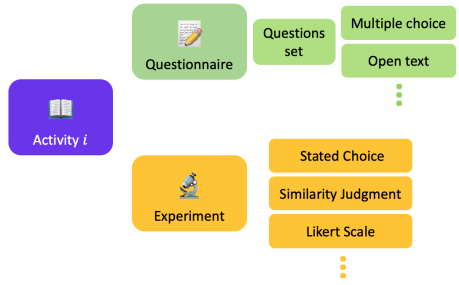



**Figure 5.3** Activities available in PixelSurvey: Experiments (Stated choice, similarity judgment, and Likert scale) and Questionnaires. Community can contribute with new experiments and questions types

Questionnaires and experiments are the two activity categories available in PixelSurvey. Within a single survey, researchers may combine multiple activities to meet their design requirements. Questionnaires consist of a sequence of one or more questions, which may include multiple-choice type, where respondents select among predefined alternatives, or open-text type, where participants provide free-text answers. Experiments, in contrast, are designed to present structured tasks based on stimuli and controlled conditions. At present, PixelSurvey implements three types of experiments: Stated Choice, Similarity Judgment, and Likert Scale (each described in detail in the following subsections). While Likert scale items are conventionally treated as a question type, in PixelSurvey they are implemented as an experiment because they can be systematically repeated across sets of images or instances, enabling controlled collection of perceptual evaluations. Although the current version provides only two question types and three experiment types, the framework has been designed to be extensible. The community can contribute additional question and experiment types, which can be integrated into future releases of the tool. For instance, people can contribute with a simple Likert question type—i.e., a single rating item without experimental repetition.



#### Closure section

The closure section contains a single page, the End page, which finalizes the survey. Its main purpose is to acknowledge the participant's contribution and provide optional debriefing information. The End page may also include a thank-you message, contact information for further questions, or a redirection link (e.g., to a participant pool system or follow-up study).

### 5.3.3 The survey recipe

As introduced, the survey recipe is the configuration unit that defines how a web survey instance is built. It is organized as a folder containing two main elements: a YAML file (the recipe file) and a source folder. These components describe the structure of the survey and provide all the resources needed for deployment.

A YAML file is a human-readable configuration document that encodes structural information in plain text. Instead of requiring users to write code, the YAML file organizes information in a hierarchical format based on key–value pairs and indentation. This makes it intuitive to read and edit, even for researchers without programming expertise. These files are also widely used in other projects such as Home Assistant, Docker, or Conda, where they serve as a simple way to define system configurations, device settings, or software environments.

#### General YAML structure

The survey YAML encodes the survey's sections and pages, the activities to be presented (e.g., questionnaires or experiments), and the parameters that control their behavior, such as experimental designs. Because it is written in a human-readable format, the YAML file allows researchers to directly inspect, modify, and share experimental designs without

requiring programming skills. This makes it possible to reproduce a survey exactly, or to adapt it by editing only a few lines of configuration. The survey YAML has the following main structure:

```yaml
survey:
  name: "My Survey"

  banner:
      title: "Title"
      subtitle: "Subtitle"
      logo: source/logo.png
      color: "#00A6D6"

  sections:
    onboarding:
      home:
        ...
      consent:
        ...
      screening:
    body:
      instructions:
        ...
      activity_1:
        ...
      activity_2:
        ...
      activity_N:
        ...
    closure:
      end:
        ...
```

The top-level key `survey` contains all information associated with a survey and is subdivided into `name`, `banner`, and `sections` properties. The `name` field specifies the project name. The `banner` field defines the settings for the top banner of the web application, including the `title`, `subtitle`, `logo`, and `color` scheme. Finally, the `sections` field declares the three mandatory components of every survey: `onboarding`, `body`, and `closure`, each containing the respective pages introduced earlier. Figure 5.4 illustrates the general layout of a survey page using the YAML banner example shown above.

Survey pages share a common layout consisting of four elements: a banner at the top, a content area, a navigation button, and a footer. The banner displays the survey title, subtitle, the user-defined logo, and the PixelSurvey logo. The main content area is usually written in Markdown format[1], but it is a dynamic space that adapts to the type of page (e.g., instructions, questions, or experiments). Markdown allows for structured text, including headings, emphasis, or lists, while remaining lightweight and easy to edit. A

[1] Markdown is a lightweight markup language that uses plain text formatting syntax, allowing for headings, emphasis, lists, and hyperlinks. It is widely used in platforms such as GitHub, Jupyter Notebooks, and many documentation systems because it is both human-readable and easy to render into HTML or PDF.


**Figure 5.4** Page layout (web) of a page in PixelSurvey

navigation button is always included to proceed to the next page (except for the End page). Finally, the footer closes the page. Specific page settings are described in detail in the following subsections.



### On-boarding/home

The Home page welcomes respondents to the survey and provides introductory information about the study. It is defined in the YAML recipe with a single property, `content`, which stores text written in Markdown format. In YAML, the "|" symbol indicates that all subsequent indented lines should be interpreted as Markdown content. This YAML snippet shows a Home page example:

```yaml
home:
  content: |
    ## Survey example
    Welcome to our survey example. In this survey you have to ...
```

### On-boarding/consent

The Consent page is structurally similar to the Home page but is designed to meet ethical requirements for online research. It contains a `content` property for displaying the consent text in Markdown format, and an additional `statement` property that defines the exact text participants must agree to before proceeding. This YAML snippet shows a Consent page example:

```yaml
consent:
  content: |
    ## Informed Consent
    I would like to invite you to participate ...
    As with any online activity, ...

  statement: "I consent to participate in this study"
```

### On-boarding/screening

The Screening page is the first point at which researchers can introduce customized questions for participants. Its purpose is to assess eligibility and, if needed, to control the composition of the sample (e.g., demographics or quota-based recruitment). The YAML definition includes the `content` property which is Markdown text to provide guidance, `questions` property to define one or more multiple-choice questions, and `quotas` to instruct the platform how to handle quota limits based on the screening question responses. This YAML snippet shows a Screening page example:

```yaml
screening:
  content: |
    ## Participant Screening
    Please answer the following questions ...

  questions:
    - id: 1
      ...
    - id: 2
      ...

  quotas:
    ...
```

The `questions` property contains a list of individual questions, each following a standard template described later in the section on survey components. These questions are multiple-choice and can be used to filter participants according to study requirements. The `quotas` property determines how responses to the screening questions are used to regulate sample composition. This allows researchers to impose limits on all possible multiple-choice alternative combinations. PixelSurvey supports three methods of quota management: *no_limit*, *uniform*, and *custom*. Figure 5.5 shows the YAML snippets for the three different configuration for quotas.

```yaml
quotas:
  method: no_limit
  fullquota_content: |
    # Thank you ...
    Unfortunately, you
    ↪  ...
```

**(a)** *no_limit*

```yaml
quotas:
  method: uniform
  limit: 100
  fullquota_content: |
    # Thank you ...
    Unfortunately, you
    ↪  ...
```

**(b)** *uniform*

```yaml
quotas:
  method: custom
  limit:
  ↪  source/quotas.csv
  fullquota_content: |
    # Thank you ...
    Unfortunately, you
    ↪  ...
```

**(c)** *custom*

**Figure 5.5** YAML configurations for the three quota modes.

PixelSurvey supports three quota management methods that regulate how respondents are admitted based on their screening responses. The *no_limit* method imposes no restrictions, allowing all participants to proceed regardless of their answers. The *uniform* method enforces balanced sampling by limiting the number of respondents for each

combination of screening responses to a predefined threshold (i.e., `limit` field). The *`custom`* method offers the highest flexibility: researchers can define quotas externally for any combination of responses, enabling complex sampling strategies such as stratified or quota-based recruitment. If a participant does not meet the eligibility criteria or if the quota for their response category has already been filled, PixelSurvey automatically terminates the survey for that participant and displays a customizable message explaining the situation (i.e., `fullquota_content` field). Table 5.1 illustrates an example of a custom quota file for two screening questions with two alternatives each, where the `quota` column defines the maximum number of respondents per bin.

**Table 5.1** Extract of a custom quota file (`quotas.csv`) for two screening questions. Each row specifies the maximum number of respondents allowed for a given combination of age and cycling frequency.

| sq1_age_label | sq2_cycling_label | quota |
|---|---|---|
| young | less than once a month | 100 |
| young | more than once a month | 200 |
| adult | less than once a month | 300 |
| adult | more than once a month | 500 |



### Body/instructions

The Instructions page is the mandatory first page of the Body section and serves as the transition between the on-boarding phase and the main research activities. The YAML definition of the Instructions page follows the same structure as the Home and Consent pages, with a single property, `content`, written in Markdown format. This is YAML snippet with a consent page example:

```yaml
instructions:
  content: |
    ## Study Instructions
    Please read these ...
```

### Body/activity

After the Instructions page, the Body section consists of one or more activities that represent the core of the survey. Each activity is declared in the YAML recipe as a separate page (e.g., `activity_1`, `activity_2`, ...), which specifies the parameters that control its behavior. PixelSurvey allows two types of activities: questionnaires and experiments. Surveys can freely combine these activities in any order to meet the requirements of a study design. The following snippet illustrates an example with two activities, the first a questionnaire and the second an experiment:

```yaml
activity_1:
  order: 1
  type: questionnaire
  persistent_instructions: |
    ## Sociodemographic questions
    Please answer the following questions ...

  questions:
    - id: 1
      ...
    - id: 2
      ...

activity_2:
  order: 2
  type: experiment
  experiment_type: ...
  persistent_instructions: |
    ## Experimental activity
    You will see pairs of ...

  task:
    ...

  settings:
    ...
```



At a general level, each activity definition includes two common properties: `order`, which establishes the sequence in which activities are presented, and `type`, which specifies whether the activity is a *`questionnaire`* or an *`experiment`*. Beyond these shared fields, each type of activity has its own structure. Questionnaires contain a single additional field, `questions`, analogous to the Screening page, where a list of multiple-choice or open-text questions can be defined. Experiments, in contrast, always require the field `experiment_type`, which specifies the type of experimental design to be implemented. They also include two additional blocks: `task`, which describes the structure of the experimental stimuli, and `settings`, which define parameters such as randomization or the number of tasks per respondent. The detailed YAML structures for questionnaires, questions, and each experiment type are described in the following subsections.

### Closure/end

The Closure section concludes the survey with a single page, the `end` page. The YAML definition contains only a `content` property written in Markdown format.

```yaml
end:
  content: |
    ## Thank you!
    We appreciate your time and contribution to this study.
```

### 5.3.4 Survey components

Activities in PixelSurvey are composed of smaller components that define their content and behavior. Questionnaires are built from a list individual questions, while experiments are defined by specific paradigms that organize tasks and stimuli. These components are encoded in the YAML recipe through structured fields, making them both transparent and extensible. In this section, we describe the currently available components, namely questions for questionnaires and three built-in experiment types. While only a limited set of components is included in the current release, the modular design of PixelSurvey enables the community to extend this library with additional question or experimental types.

#### Questions for questionnaires

Questions are the fundamental elements of questionnaire activities. Each questionnaire is composed of a list of questions, with each entry in YAML corresponding to a single item. A question element always requires `id`, `type`, `order`, and `question`. Depending on the type, additional properties specify the available responses or constraints. At present, PixelSurvey supports two question types: `multiple-choice` and `open-text`.



**Multiple-choice questions** Multiple-choice questions allow participants to select one predefined alternative from a set of options. In PixelSurvey, these are rendered as drop-down menus, but they can also be adapted to other interface styles such as radio buttons or check boxes with community-contributions. Multiple-choice questions can appear either in the Screening page or in a Questionnaire activity, always defined under the `questions` field. This is a YAML example for a multiple-choice question:

```yaml
- id: 1
  type: multiple_choice
  order: 1
  question: "What is your age group?"
  alternatives:
    - "18-25"
    - "26-35"
    - "36-45"
    - "46+"
  required: true
  variable_name: age_group
```

In this format, the `question` field specifies the text presented to participants, and `alternatives` lists the set of available response options. The `required` field specifies whether an answer must be provided before proceeding; it must be set to *`true`* in Screening pages but can be freely defined in questionnaires. Finally, the `variable_name` field defines the identifier under which responses are stored in the database, ensuring that data can be consistently retrieved and analyzed.

**Open-text questions** Open-text questions provide participants with a free-form input field to enter qualitative responses. This type is particularly useful for capturing opinions, justifications, or information that cannot be predefined as a set of alternatives. Unlike multiple-choice items, open-text questions do not include additional response fields and can only be used in questionnaire activities, not in the Screening page. The YAML definition for an open-text question is shown below:

```yaml
- id: 2
  type: open_text
  order: 2
  question: "What factors influence your choice of transport mode?"
  required: false
  variable_name: transport_factors
```

Multiple-choice and open-text are basic forms of questions useful for most survey designs. The system, however, is extensible: additional formats—such as sliders, ranking tasks, Likert scales, or image-based inputs—can be integrated in future versions through community contributions.



### Experiment types

Experiments in PixelSurvey are designed to implement established paradigms in behavioral research. They consist of structured tasks where participants evaluate or choose between stimuli under controlled conditions. Each experiment includes a `task` block, describing what participants see and how they should respond, and a `settings` block, which defines how tasks are generated (e.g., randomization, number of repetitions). PixelSurvey currently implements three experiment types: Stated Choice, Similarity Judgment, and Likert Scale.

The selection of these three experiment types for the initial release of PixelSurvey is motivated by their relevance in behavioral research and their methodological complementarity among them. Stated Choice represents a dominant paradigm in transportation, marketing, and economics for analyzing trade-offs and estimating preferences. Similarity Judgment, in turn, captures perceptual organization and comparative reasoning, and has a long tradition in psychology and cognitive science [176, 178]. Both paradigms are particularly well-suited to image-based experiments, where visual stimuli form a central part of the evaluation. Finally, the Likert Scale experiment offers a flexible paradigm for obtaining repeated evaluations of images on key dimensions, which can also complement the data generated by the other two experiments. The following subsections describe each experiment type and its implementation in PixelSurvey.

**Stated Choice (SC) experiment** Stated choice (SC) experiments ask respondents to select one option among two or more alternatives defined by attributes [156]. They are widely applied across fields such as transportation, marketing, environmental and health economics, and consumer research. Their primary purpose is to generate data for

estimating discrete choice models [188], which provide insights into economic behavior by quantifying trade-offs and identifying the relative importance of different attributes. In doing so, SC experiments allow researchers to infer measures such as willingness-to-pay.

PixelSurvey implements a binary SC experiment to trade off between two alternatives. One attribute of the alternatives can be an image. This makes it possible to combine traditional textual or numerical attributes (e.g., price, size, travel time) with visual information (e.g., facade design, landscape features) [178]. The following YAML snippet describes an SC experiment, where `experiment_type` is set to *`stated_choice`*:

```yaml
activity_1:
  order: 1
  type: experiment
  experiment_type: stated_choice
  persistent_instructions: |
    ## Activity 1: Choosing your right house
    In the following ...

  task:
    query: "Which option do you prefer?"
    instance: "House"
    n_alternatives: 2
    attributes:
      - label: "Facade"
        type: image
        unit: "url"
      - label: "Price"
        type: standard
        unit: "€"
      - label: "Rooms"
        type: standard
        unit: ""

  settings:
    ...
```

The YAML example defines a stated choice experiment with two alternatives. The `task` block specifies the essential elements of the experimental task. The `query` field provides the question presented to participants, the `instance` field labels the type of object being compared (in this case, “House”), and the `n_alternatives` field indicates the number of options shown per task. At present, PixelSurvey supports choice sets with only two alternatives. Attributes for each alternative are then listed individually. Each attribute includes a `label`, which is the textual representation of the attribute; a `type`, which can be *`image`* for visual attributes or *`standard`* for numerical or textual ones; and a `unit`, which is set to *`url`* for images but can otherwise be freely defined. In this example, the facade of the house is represented as an image attribute, while price and number of rooms are defined as standard numerical or textual attributes. Figure 5.6 shows the layout that this YAML SC configuration generates.

Finally, the `settings` block controls how choice sets are generated and presented to participants. Currently, two different settings modes can be used for all experiments: *`custom`* or *`random`*. Figure 5.7 shows the YAML snippets for both modes.

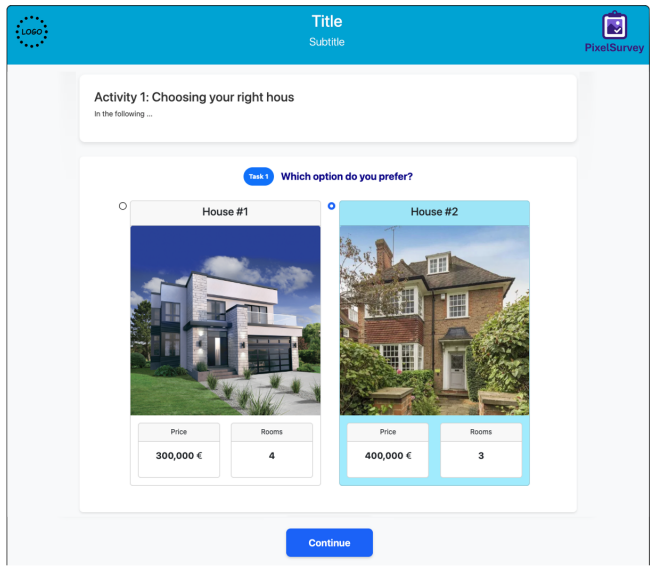



**Figure 5.6** Layout generated with the SC experiment YAML presented

The `settings` block determines how experimental tasks are generated and assigned to respondents. The field `tasks_per_respondent` specifies the number of choice tasks that each participant must complete. The `experimental_design_mode` can be set to either *`custom`* or *`random`*. In the custom mode, the researcher supplies a pre-specified experimental design through an external CSV file (`custom_design`), allowing full control over the combinations of attribute levels shown to respondents. Table 5.2 illustrates the content of a custom design file corresponding to the YAML example. Each row represents a task with a unique `task_id`, which is grouped under a `set_id`. All tasks with the same `set_id` are completed by the same respondent. Each column specifies the attributes of the alternatives to be presented.

**Table 5.2** Excerpt of a `custom_design` file used for full specification of SC tasks.

| task_id | set_id | alt1_att1_facade | alt1_att2_price | alt1_att3_rooms | alt2_att1_... |
|---|---|---|---|---|---|
| 1 | 1 | http://.../img8.png | 100000 | 2 | ... |
| 2 | 1 | http://.../img5.png | 100000 | 1 | ... |
| 3 | 1 | http://.../img3.png | 300000 | 2 | ... |

In the random mode, attribute values are drawn randomly from a predefined CSV file (`attributes_values`). The number of generated sets is defined by the field `number_of_sets`. Table 5.3 shows an example of an attributes values file used to generate random choice sets. The `attribute_name` column specifies the attribute listed in `task/attributes`, and the `attribute_value` column enumerates all possible values used to generate random combinations.

```yaml
settings:
  tasks_per_respondent: 8
  experimental_design_mode: custom
  custom_design:
  ↪  source/custom_design.csv
```

**(a)** *custom*

```yaml
settings:
  tasks_per_respondent: 5
  experimental_design_mode: random
  number_of_sets: 100
  attributes_values:
  ↪  source/att_values.csv
```

**(b)** *random*

**Figure 5.7** YAML configurations for the two settings modes in SC experiments.

**Table 5.3** Excerpt of an `attributes_values` CSV file used for random generation of SC tasks.

| attribute_name | attribute_value |
|---|---|
| facade_url | http://www.example.com/img1.png |
| facade_url | http://www.example.com/img2.png |
| facade_url | http://www.example.com/img3.png |
| price | 100000 |
| price | 200000 |
| price | 300000 |
| rooms | 1 |
| rooms | 2 |



**Similarity Judgment (SJ) experiment** Similarity judgment (SJ) experiments ask respondents to evaluate the perceptual similarity or dissimilarity among instances. A common task is the "odd-one-out", where participants identify which item in a triplet is most different compared to the other two. These designs are widely used in psychology and cognitive science to study perceptual organization, category formation, and conceptual reasoning [189]. More recently, they have been employed in computational cognitive science and machine learning to collect large-scale perceptual data and build similarity-based embeddings [175, 176]. Their main purpose is to elicit structured judgments that reflect how humans perceive and organize stimuli, which can then be used to validate theories of perception, construct cognitive models, or align computational representations with human intuitions.

PixelSurvey implements an SJ experiment that compares sets of instances, where each instance is composed of one or more images. The following YAML snippet defines an SJ experiment, where `experiment_type` is set to *`similarity_judgment`*:

```yaml
activity_2:
  order: 2
  type: experiment
  experiment_type: similarity_judgment
  persistent_instructions: |
    ## Activity 2: Which is the odd...
    In the following ...

  task:
    query: "Which house interior is the odd one out?"
    instance: "House interior"
    images_per_instance: 3

  settings:
    ...
```

In this example, each instance corresponds to a "House interior". The `task` block specifies the `query` field, which defines the question presented to participants; the `instance` field, which labels the type of object being compared; and the `images_per_instance` field, which indicates how many images are used to represent each instance (three in this case). Figure 5.8 shows the layout that this YAML SJ configuration generates.



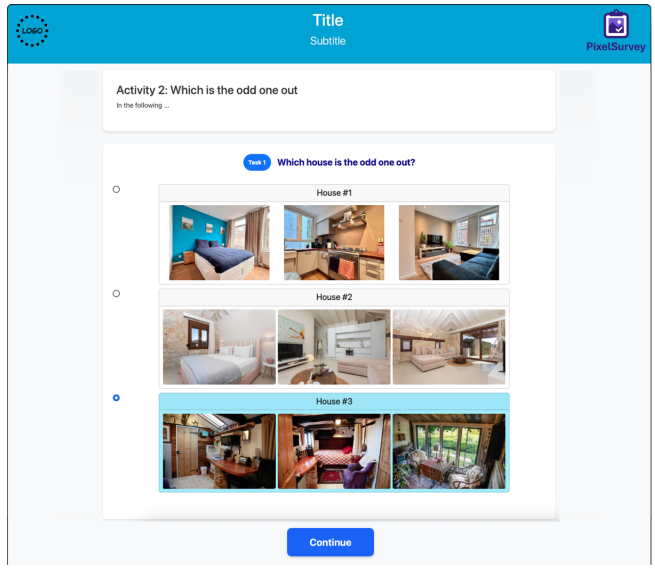



**Figure 5.8** Layout generated with the SJ experiment YAML presented

The `settings` block is analogous to the SC experiment and controls how task sets are generated and presented. Figure 5.9 shows the YAML snippets for *`custom`* and *`random`* modes.

```yaml
settings:
  tasks_per_respondent: 8
  experimental_design_mode: custom
  custom_design:
  ↪  source/custom_design.csv
```

**(a)** *custom*

```yaml
settings:
  tasks_per_respondent: 5
  experimental_design_mode: random
  number_of_sets: 100
  instances: source/sj_instances.csv
```

**(b)** *random*

**Figure 5.9** YAML configurations for the two settings modes in SJ experiments.

In *custom* mode, the researcher provides a pre-specified experimental design through an external CSV file (`custom_design`), formatted similarly to SC but with columns that directly list the images associated with each instance. In *random* mode, the key difference with SC that an `instances` CSV file is required instead of an `attributes_values` file. Each row in the `instances` file defines a complete instance, including an identifier and the corresponding image URLs.



**Likert Scale (LS) experiment** Likert scale (LS) experiments ask participants to evaluate one or more dimensions of an instance on a predefined ordinal rating scale (e.g., from 1 to 5) [190]. Likert scaling has become one of the most widely used tools for measuring attitudes, perceptions, and evaluations across disciplines such as psychology, sociology, education, marketing, and health research [191]. Its primary strength lies in capturing subjective judgments in a standardized, repeatable way that facilitates both descriptive analysis and the construction of latent constructs (e.g., attitudes or satisfaction indices). Unlike a traditional Likert-style survey question, which can be potentially developed as a single item in a questionnaire, the LS "experiment" is designed as a structured task that can be repeated across multiple instances and conditions.

Currently, PixelSurvey implements LS as experiments, not as a single-question format. This format allows researchers to systematically measure perceptions, attitudes, or evaluations under controlled designs. The following YAML snippet defines an LS experiment, where `experiment_type` is set to *likert_scale*:

```yaml
activity_3:
  order: 3
  type: experiment
  experiment_type: likert_scale
  persistent_instructions: |
    ## Activity 3: Rating task
    Please rate the following criteria ...

  task:
    query: "Rate the following dimension about the house"
    instance: "House"
    likert_scale: 5
    evaluations:
    - label: "Safety"
    - label: "Aesthetics"

  settings:
    ...
```

In this example, the `task` block defines the rating activity. The `query` field specifies the statement or prompt shown to participants, `instance` labels the type of object being evaluated, `likert_scale` sets the number of points on the rating scale, and `evaluations` lists the specific dimensions to be rated. Figure 5.10 shows the layout that this YAML LS configuration generates.

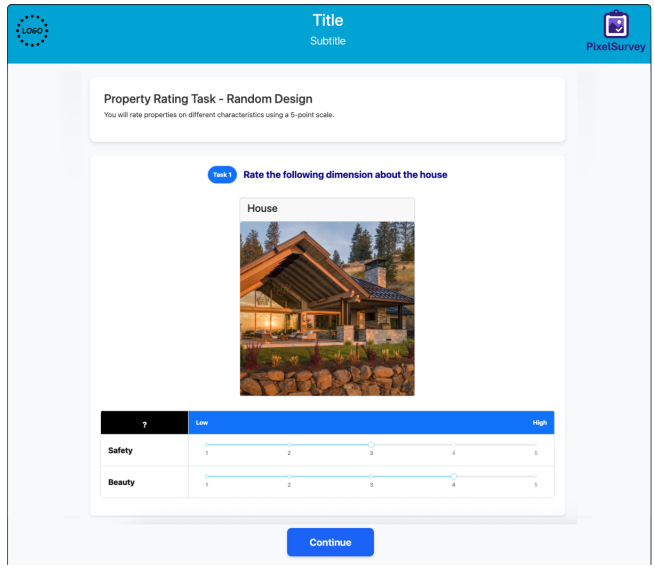



**Figure 5.10** Layout generated with the LS experiment YAML presented

The `settings` block follows the same structure as in the SJ experiments. In *`custom`* mode, a full design is provided through a `custom_design` CSV file, where each row corresponds to a task and set, specifying the instance to be shown and rated. In *`random`* mode, tasks are generated from instances listed in an external CSV file (`instances`), where each row corresponds to a single instance, which is represented by an image URL.

### 5.3.5 PixelSurvey Output

Once the YAML recipe and source files are provided to **PixelSurveyCore**, the framework automatically generates a dedicated `Survey` folder. This folder contains all the elements required to execute the survey as a web application and to store participant responses. In other words, the output of PixelSurvey is a web app with an embedded database for managing experimental data and responses. Figure 5.11 shows the full folder structure for an output survey.

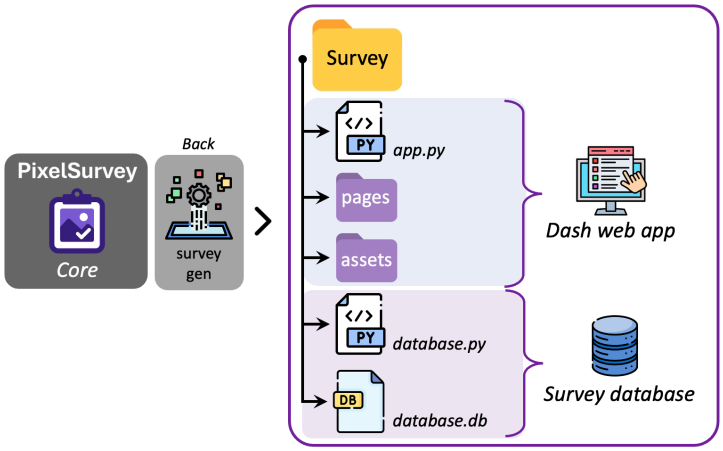



**Figure 5.11** PixelSurvey output: the framework generates a `Survey` folder containing a Dash web app (top) and a dedicated survey database (bottom).

#### Dash web application

The first component of the output is a Dash-based web application coded in Python, which participants interact with to complete the survey. This application is defined by a set of automatically generated files:

- `app.py`, which provides the entry point of the application.
- `pages/`, a folder containing individual pages of the survey (e.g., screening, and the different activities).
- `assets/`, which contains static resources such as logos or stylesheets.

These files form a self-contained web service that can be directly deployed to collect responses. Because all pages are created from the YAML recipe, the generated app exactly reflects the survey design specified by the researcher, ensuring reproducibility and transparency.

**Survey database**

In parallel, PixelSurvey automatically creates a dedicated database to manage experimental tasks and responses. The database is implemented in SQLite format and is stored in the `Survey` folder as `database.db`, with its manager defined `database.py`. The *db* file stores the different tables for generating the experiments and collecting responses, and the *py* file implement the methods for interaction between the web app and the database. This means that researchers do not need to manually configure or maintain the database; instead, they can directly query it to analyze collected data. The internal structure of the database tables is detailed in the following subsections.



**Quotas table** The quotas table records the distribution of respondents across different screening categories and monitors quota fulfillment. It mirrors the structure defined in the screening page (see Table 5.1), with an additional column that tracks how many participants from each combination of screening responses have already completed the survey. This allows the system to enforce the quota rules in real time as new participants attempt to enter the study.

**Experimental task table** The experimental task table defines the pool of tasks included in each experimental activity. A separate table is created for every experiment declared in the YAML recipe. Each table begins with a unique `task_id` column, followed by the fields required by the corresponding experiment type. For example, in stated choice experiments, `task_id` is followed by the attribute values of two alternatives that together define a single choice task.

**Task set table** The task set table organizes experimental tasks into sequences that are assigned to individual respondents. Each row corresponds to a unique `set_id`, which groups together the `task_ids` that a participant must complete. For example, if a survey contains two experimental activities—one with three tasks and another with four—each row of the Task set table will include a `set_id` and seven corresponding `task_ids`, three from the first activity and four from the second.

**Response table** The response table stores all participant responses collected during the survey. Each row corresponds to a `respondent_id` and includes answers to screening questions, questionnaire individual questions, and experimental tasks. Each record links the response to the corresponding participant, activity, and task, ensuring that the full trajectory of individual responses can be reconstructed.

**Timestamp table** The timestamp table logs the precise moment, in microseconds, when participants click the navigation button to move between pages. This provides detailed information about response times at the page level (i.e., task level), supporting analyses of task duration and participant engagement.

## 5.4 Usage

This section outlines the typical workflow for running a study with PixelSurvey: accessing PixelSurveyCore, preparing a survey recipe, generating the application, deploying it, collecting data, and exporting responses.

### 5.4.1 Step 1: Accessing PixelSurveyCore

PixelSurveyCore is developed in Python and builds survey applications on top of the Plotly Dash framework. The source code is openly available on GitHub and can be cloned or downloaded directly. Once installed, researchers can generate surveys by providing a recipe (YAML file and source folder) as input.

In addition to the command-line version, a lightweight graphical user interface (GUI) is available. In this variant, users can upload a compressed survey recipe (ZIP file), and the platform automatically compiles and returns the corresponding web application. This lowers the entry barrier for non-technical users.



### 5.4.2 Step 2: Creating a survey recipe

The second step is to define the survey recipe. As described in the previous section, this consists of a YAML file and a source folder. The YAML file encodes the structure of the survey, including its sections, activities, and settings. The source folder contains all supporting files referenced in the YAML, such as logos, optional quota definitions, and experimental designs in CSV format.

### 5.4.3 Step 3: Generating the web application

Once the recipe is prepared, the next step is to generate the survey web application. In the command-line version of PixelSurveyCore, this is done by running the `survey_gen` command that compiles the YAML file and source folder into a ready-to-use web application. Internally, the system processes the recipe, creates the necessary database, assembles the front-end and back-end components of the survey, and generates a Dash App project ready to be deployed.

In the GUI version, users upload the ZIP file containing the recipe, and the platform compiles it. At this point, users can either directly download the generated Dash application for local or institutional deployment, or opt to host it directly on the PixelSurvey server. In both cases, the result is a self-contained survey application ready for immediate use.

### 5.4.4 Step 4: Deployment and data collection

Once the app is generated, the survey application can be deployed. Researchers have two options: they may download the Dash application and run it locally or on institutional servers, or they may choose to host it directly on the PixelSurvey platform. Hosting on the platform simplifies the deployment and allows surveys to be launched with minimal technical setup, while local deployment provides full control.

There are some trade-offs between local and PixelSurvey-hosted deployment. Hosting locally or on institutional servers gives researchers full control over data storage, access, and compliance with institutional or legal requirements (e.g., GDPR). This approach is generally recommended for large-scale studies, sensitive data, or projects that require strict security policies. By contrast, hosting on the PixelSurvey server lowers technical barriers and simplifies deployment, making it suitable for pilots, classroom use, or small/medium-scale studies. However, this option requires trust in the platform's infrastructure for handling and storing data, and may be less suited to sensitive or large-scale deployments. Researchers should evaluate their technical resources, ethical obligations, and study requirements before deciding on a hosting option.



For local or institutional deployment, the GitHub repository includes a ready-to-use `docker-compose` recipe that streamlines the process of setting up the server environment. Docker is a lightweight containerization system that packages software together with all its dependencies. This ensures that applications run reliably on the same context. Running the provided `docker-compose` file automatically configures the required services, offering a straightforward way to get the survey system operational without additional manual setup.

Researchers with little or no experience in server administration can request assistance from their institutional IT departments for deployment. In practice, the requirements are minimal: the IT team only needs to set up a server with web ports 80 (HTTP) and 443 (HTTPS) open, install Docker, and run the provided `docker-compose` file together with the output survey folder. This setup automatically builds the PixelSurvey application and database without requiring manual configuration. In future releases, PixelSurvey aims to provide a simplified deployment script that prepares the server environment even more transparently, further easing the workload for IT staff and lowering the barrier for researchers to run their own studies.

During data collection, all participant interactions are stored automatically in the survey's SQLite database where the survey application is running.

### 5.4.5 Step 5: Exporting and accessing data

After data collection, responses are stored in the SQLite database automatically generated with each survey instance. The database contains all relevant tables, including screening responses, questionnaires, experimental tasks, and timestamps. Because it is implemented in SQLite, the file can be easily accessed using a wide range of tools. For example, researchers can open the database directly in Python using `pandas` or in R through packages such as `RSQLite`. Alternatively, any standard SQLite database manager can

be used for inspection. Data can be queried directly from the database or exported to CSV format for further statistical analysis, modeling, visualization, or integration with external workflows.

## 5.5 Open-science and community contributions

PixelSurvey has been created as a survey tool that embeds open-science principles into its core. Its architecture emphasizes transparency, reproducibility, and collaborative extension, making survey designs easier to share, replicate, and build upon. The subsections below describe how this philosophy is implemented and how researchers can contribute to the framework's continued development.

### 5.5.1 Open-science philosophy

PixelSurvey has been designed as a tool which embraces and advances open science. Its reliance on human-readable YAML recipe files ensures that survey designs are easy to create, and straightforward to share. With PixelSurvey it is easy to share survey logic, structure, and stimuli considered. Additionally, its framework allows transparency and reproducibility, as other researchers can directly re-run, adapt, or extend a study simply by re-using the original recipe file.



PixelSurvey incorporates explicit versioning practices to ensure long-term reproducibility. Each YAML recipe includes a version number key, which instructs the PixelSurveyCore compiler to interpret the file with the appropriate specification. This ensures that older YAML recipes remain functional even as the platform evolves, making them durable scientific artifacts comparable to datasets, experimental designs, or analysis scripts. Recipes can therefore be archived, cited, and re-used alongside publications, supporting both replication and cumulative research.

### 5.5.2 Community contributions

The open-source and modular design of PixelSurvey encourages community contributions that extend its functionality. Contributions can take the form of reusable components. This version of PixelSurvey allows the contribution of additional questionnaire item types or entirely new experimental paradigms. Integration into the framework occurs through pull requests to the PixelSurveyCore GitHub repository, ensuring transparent review and collaborative improvement. To contribute a new component, developers are asked to provide the following elements:

- **YAML definition**: A snippet specifying the configuration of the component, following the conventions illustrated for existing question types (e.g., multiple-choice, open-text) and experiments (e.g., stated choice, similarity judgment, Likert scale). The YAML must declare all required structural elements.

- **Parser file**: A Python class that maps the YAML configuration into internal variables accessible within the framework. This file ensures that the declared keys and values are correctly processed by the system.
- **Front-end component**: A Python class implementing the user interface of the component, built on Plotly Dash [192] and Dash Bootstrap Components [193]. All components are organized as `Card` elements, which can contain standard web widgets such as text boxes, radio buttons, or sliders.
- **Source file structure (for experiments)**: A clear definition of any external files required to instantiate experimental tasks. These are the CSV files containing attributes, or instances. This guarantees that the platform can correctly load, display, and randomize tasks.

Looking forward, we envision PixelSurvey as a continuously evolving ecosystem. Planned developments include the addition of a graphical user interface (GUI) to lower entry barriers for non-technical users, the integration of further experimental paradigms, and community-driven libraries of survey templates. In this way, PixelSurvey aims not only to support current research needs but also to anticipate future demands by combining open science principles with collaborative software development.



## 5.6 Conclusions

PixelSurvey was developed to fill the absence of a flexible, open-source framework capable of supporting experimental designs with image-based stimuli at scale. It enables the creation of survey experiments in which large image datasets can be easily integrated into experimental designs. The framework lowers technical barriers by adopting a modular architecture based on human-readable YAML files. These files also ensure that surveys can be shared, archived, and replicated as scientific artifacts, thereby contributing directly to open-science practices.

The framework is able to create surveys with two question types and three experiment paradigms — i.e. Stated Choice, Similarity Judgment, and Likert Scale. Those experiments cover a wide spectrum of behavioral research needs. Its extensibility, however, ensures that new components can be further expanded over time, whether developed by the core team or contributed by the community. Through its design, PixelSurvey broadens methodological possibilities for image-based studies and provides a robust platform for creating experiments built on textual or numerical inputs.

Looking ahead, the continued development of PixelSurvey will be driven by community engagement and the integration of additional experimental paradigms, user-friendly interfaces, and shared libraries of survey templates. In doing so, the framework aspires to become a long-term resource for researchers across disciplines, fostering cumulative knowledge building and accelerating innovation in the study of human perception, preferences, and decision-making.

# Chapter 6

# From pixels to perceptions


## Abstract

This study advances perception-aware urban representation learning by integrating human similarity judgments into computer vision models. This model aims to integrate the components, captured by the pre-trained computer vision model, with perceptions, captured by the similarity judgments. Participants evaluated triads of urban scenes, selecting the image most distinct from the others. These perceptual relationships were then used to supervise the training of an embedding model, aligning machine-learned visual representations with human perceptual structure. The resulting "perception-aligned" embeddings capture finer distinctions in how people group and differentiate urban scenes, beyond what unsupervised models can achieve. Conceptually, this study bridges the physical (component) and perceptual (condition) layers of urban imagery, demonstrating how human cognition can guide computer vision models to better represent human perceptions. Specifically, this uses do not use the concept of condition, but it works with both kind of information. It establishes the methodological and conceptual transition of the thesis—from automated component extraction toward the integration of human perception in urban visual analytics.


This chapter is based on the journal article: Garrido-Valenzuela, F., Cats, O., & van Cranenburgh, S. (2025). *From pixels to perceptions: Using human similarity judgments to enrich urban-space embeddings*. International Journal of Geographical Information Science. Code and Data are available at the repository: Garrido Valenzuela, Francisco; Oded Cats; Sander van Cranenburgh (2025): Code underlying the publication: From pixels to perceptions: using human similarity judgments to enrich urban space embeddings. 4TU.ResearchData.

## 6.1 Introduction

Multidimensional spatial urban data play an increasingly important role in shaping the quality of life in cities. In an era where cities are becoming more complex and dynamic, these data enable strategic urban planning [194] and support evidence-based decision-making [195]. Multidimensional spatial data refers to geographical datasets that comprise various aspects of urban areas, including but not limited to infrastructure, environment, demographics, and behavioral dimensions. With this set of dimensions, city planners can holistically identify areas for potential development [196], optimize resource allocation [197], and address demanding urban challenges [198, 199]. For instance, if planners want to foster walkability, they can identify neighborhoods with inferior pedestrian infrastructure and high car dependency. This insight allows them to further investigate the properties of these areas and promote safer and more walkable areas through the implementation of pedestrian-friendly initiatives, such as widened sidewalks, crosswalk enhancements, and traffic calming measures. Overall, integrating multidimensional spatial data into urban planning supports policymakers in making informed decisions that promote social, economic, and environmental well-being.

Recently, advances in representation learning have made considerable progress in producing high-quality and multidimensional spatial data by combining different sources. Scientists in this field have developed machine learning models to combine distinct types of data into vector representations (aka embeddings), most commonly text or images [200–202]. Such embeddings are low-dimensional representations of the original data but preserve their meaningful characteristics and associations. Specifically in the urban computing field, researchers have extensively used embeddings produced from different geo-tagged data sources while preserving associations with city-related attributes [43, 45, 46, 203]. Urban Space Embedding Models (USEM) can learn and produce urban representations (i.e., vectors) from a combination of data sources such as points-of-interest (POIs) locations [204], travel demand flows from GPS traces [205], and visual content from street-level images (SLI) [206] or satellite images. Subsequently, these vectors can be used by planners or policymakers to explore the structure of cities and understand different urban phenomena. For example, urban embeddings can be used to identify the number of neighborhood types and their spatial distribution, to analyze similarities across urban areas, or to explore associations among different urban characteristics.

Evidence from urban planning studies in psychology and sociology highlights the significance of human perceptions, such as safety, attractiveness, or vibrancy, when it comes to evaluating and experiencing urban spaces [207]. At the city scale, crowd-sourced studies that leverage street-level images have shown how visual cues can be used to map perceived safety, beauty, and socio-economic conditions [22, 47, 208], and a comprehensive overview of visual-intelligence approaches is provided by [209]. These subjective aspects are relevant for policy making, as they provide people-centered insight into the design of public spaces and urban policies [51]. For instance, perceived safety in neighborhoods affects people's mental well-being [210] and correlates with the amount

of physical activity [211, 212]. In addition, [75] demonstrate that the perceived quality of urban spaces can promote social interactions within these spaces. Urban perceptions can assist urban planners and researchers in understanding how humans use urban spaces [213] and consequently play an instrumental role in designing more livable places.

Despite the advancements in representation learning techniques, existing models often face challenges in incorporating human perceptions into urban embeddings. Human perceptions are intrinsically subjective and context-dependent, making them difficult to measure and quantify using traditional computational approaches. For example, platforms commonly used for crowd-sourcing (e.g., [214]) struggle in controlling the demographics and cultural backgrounds of participants. In addition, people's perceptions of the same place may differ. For instance, someone who usually walks may have a different perception of the greenness in a neighborhood compared to someone who usually drives. So, incorporating perceptions in the loop supports the design of inclusive and people-centered urban interventions by considering a diverse set of preferences in relation to the design of urban spaces. The subjective nature of these perceptions and the difficulty in measuring them consistently across diverse populations contribute to a scarcity of quantified perception data [22]. Furthermore, capturing human perceptions regarding urban spaces requires a deep understanding of cultural, social, and psychological factors, which may not be fully captured by machine learning algorithms alone. As a result, current urban embedding models do not incorporate people's perceptions to accurately represent residents' experiences into the urban vectors. Thus, there remains a significant knowledge gap in existing approaches to urban embedding, which fail to account for how humans perceive urban space.



To address this gap, we propose a method for constructing and training an USEM using SLI and incorporating human perceptions of public urban spaces. Our approach is inspired by human similarity judgments from behavioral and cognitive science [175]. Similarity judgments are comparisons and evaluations people can make about the likeness between different entities based on objective and perceived characteristics. Specifically, we conducted a human similarity experiment where participants compared different urban spaces using SLI. Psychological experiments suggest that when people compare places, such judgments inherently involve trade-offs among various attributes of the places [215, 216]. These attributes may include objective factors, such as the amount of vegetation, number of cars, or architectural style, as well as subjective perceptions, such as safety, beauty, or vibrancy. The collected judgments allow the USEM to learn how similar or different urban spaces are based on these metrics. The perceived similarity reported by people is therefore used as a guiding principle for training the model, thereby allowing us to effectively capture visual and perceptual information about the urban images to include them in the embedding model.

The main contribution of this study is the incorporation of human perceptions into machine-learned representations of urban space through computer vision. We introduce an approach for learning general-purpose perceptual urban space embeddings that integrates both visual features and subjective human judgments. These dimensions are not pre-specified and emerge jointly from the images and human responses. Our

proposed method consists of three main steps. First, we define the spatial unit of analysis using polygonal areas and associated geo-tagged SLI. Second, we conduct a similarity judgment experiment, where participants compare triplets of SLI and select the spatial unit that stands out most from the others based on their personal perception. These human responses serve as a similarity signal for training our model. In the third step, we train a deep metric learning model using these triplet comparisons, enabling the USEM to encode not only visible neighborhood components such as building types, street furniture, and infrastructure but also latent perceptual elements informed by how people visually interpret urban environments.

We apply this method within the Netherlands to demonstrate its effectiveness. Specifically, we draw around one million geotagged-SLI from across the country using the approach described by [132]. The collected images were used to execute the similarity experiment within people living in the country, and then used for training the USEM. Once we have trained the model, it is applied in Rotterdam, the second largest city in the country, to showcase different perceived urban areas captured by our model and its applications.

The remaining parts of this document are organized as follows. In Section 6.2, we present an overview of urban embedding models and human similarity judgments. Then, we describe the input data required for implementing our method and how we collected the data for the Netherlands. Section 6.4 describes the method and showcases its implementation in the Netherlands. Finally, the results of the model's application in Rotterdam, the discussion and the main conclusions are reported.



## 6.2 Related work

This section provides an overview of the concepts and methods relevant to this study, focusing on urban representation learning and human similarity judgments. First, we review state-of-the-art USEMs, highlighting the diverse approaches and data sources involved in their development. Next, we examine the theories behind human similarity judgments, commonly used in behavioral and cognitive science to understand the structure of individuals' mental representations of their environment. By integrating these two domains, this study aims to incorporate human perceptions into urban embedding models, producing multidimensional spatial data that reflects these perceptions.

### 6.2.1 Urban representation learning

Urban representation learning lies at the intersection of urban computing and metric learning, focusing on representing urban regions as vectors. Similar to [3], who introduced word embeddings (i.e., Word2Vec) for representing words as vectors, urban representation learning uses metric learning to automatically derive a representation function that maps urban regions into an embedded space. The distance in the embedded space should

preserve the regions' similarity, ensuring that similar urban regions are positioned close to each other while dissimilar regions are distanced. Figure 6.1 conceptualizes the process and goal of creating an urban embedding.

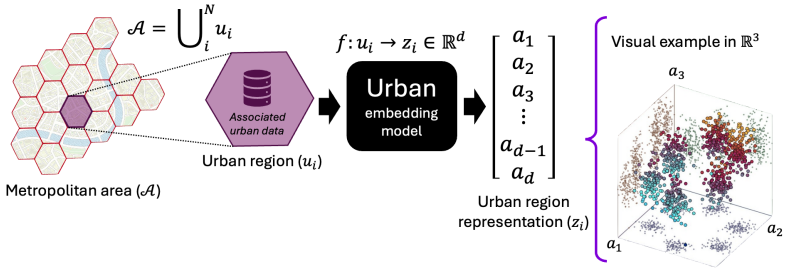



**Figure 6.1** General procedure for creating an urban space embedding in the urban representation field. A metropolitan area $\mathscr{A}$ is subdivided into a set of urban regions $u_i$. All data in $u_i$ is processed and mapped into $\mathbb{R}^d$. A visual example in $\mathbb{R}^3$ is shown. Notation is based on [45]

This process is formalized as the *Urban Neighborhood Embedding Problem* [45], which states that a metropolitan area $\mathscr{A}$ is composed of a set of urban regions $u_i$. Each region may contain different urban data such as street-level images, points of interest (POI) and/or travel flows. Then, a model has to learn a function $f$ that maps each $u_i$ into $\mathbb{R}^d$. As a result, this model transforms and combines multidimensional characteristics of urban environments (i.e., input data) into compact representations. This facilitates the understanding of urban areas and enables various applications in urban planning and analysis. For instance, these compact representations can be used to identify and classify neighborhood types, analyze urban mobility patterns, assess the impact of urban interventions, predict socioeconomic indicators in data-limited areas, or support location-based services and recommendations.



### Urban region similarity

In the urban field, various approaches have been proposed to measure urban similarity, with the most common practice relying on Tobler's First Law of Geography [217]. This law states that "everything is related to everything else, but near things are more related than distant things", implying that spatial proximity can serve as a proxy for urban region similarity. The key principle for training an urban embedding model is to establish a predefined logic to determine similarity among urban regions. For instance, in Natural Language Processing (NLP), two words are considered similar if they share a similar set of surrounding words in context. By applying this logic, models can learn patterns that define word similarity. Similarly, for urban embedding models, defining a metric for similarity, such as spatial proximity or other characteristics, allows the model to learn and capture the nuanced similarities among urban regions.

For instance, the urban embedding model Hex2Vec [46] uses Tobler's law for sampling urban regions and employs the Skip-gram model with negative sampling [3]. The Skip-gram model, initially designed for word embeddings, requires positive (i.e., similar instances) and negative samples (i.e., dissimilar instances) for each region in the training set. Adjacent regions to a target one can serve as positives, while distant ones can serve as negatives. Similarly, Loc2Vec [218], Tile2Vec [219], and Urban2Vec [43] use Tobler's law for sampling regional instances. These models sampled triplets of regions for use with the triplet margin loss [31]. This loss function enables the model to learn similarity constraints based on one positive and one negative instance for each anchor region. Also, triplet loss tends to capture information more efficiently compared to other loss functions, such as pairwise loss [220]. Alternative approaches to urban embedding do not rely directly on Tobler's Law. For instance, the RegionEncoder model [205] uses taxi GPS traces to define sequences of regions. These sequences are treated as similar instances, assuming that regions frequently visited in succession share functional or contextual similarities. This method emphasizes the functional connectivity between regions rather than their spatial proximity.

**Data sources used in urban embeddings**

Various data sources have been utilized to create representations of urban spaces, offering insights into domains such as transportation, infrastructure, urban amenities, and human behavior. Commonly used datasets include POIs, which provide the location of different activities, services, and infrastructure within an area; satellite images, which offer high-resolution views of the physical layout and land use patterns; street-level images, which provides visual context of the surroundings; and mobility traces, such as GPS from taxis or mobile devices, which provide movement patterns and network connectivity. While initial approaches focused on using one type of data, recent research in this field emphasizes the development of multi-modal embeddings, which combine various datasets [45].

Integrating diverse sources of information, such as POIs, satellite images, street-level images, and mobility traces, these models can produce a more complete representation of urban environments for a wider range of applications. For instance, [221] utilized OpenStreetMap (OSM) building footprints combined with points of interest (POIs) to create urban space embeddings. Their work captures the spatial and functional characteristics of urban areas. Similarly, [222] combined satellite images with POIs, adding a bird's eye visual component to the function provided by the amenities. Urban2Vec [43] integrated multi-modal data, including Street View images and textual data related to POIs, to enhance the contextual understanding of urban neighborhoods. Additionally, [45] proposed a general multi-modal framework for using any geo-tagged data within urban regions together with the road network to build an urban space embedding.

### 6.2.2 Human similarity judgments

Human similarity judgments refer to how people assess the likeness or difference between entities (or concepts). This involves comparing physical features such as color and shape, as well as abstract and perceived characteristics like category or function [223]. For instance, a knife can be perceived as a weapon with a dangerous connotation, or as a utensil with a practical connotation [216]. These judgments are grounded in cognitive processes that evaluate similarities and differences among various stimuli based on human perceptions and experiences [224]. They can therefore vary based on context, personal experience, and cultural background. For example, some people may think a cat is more similar to a dog compared to a tiger because both are pets, and humans are closer to them, while others may think that a cat is more similar to a tiger, considering that both are felines. This variation in judgments highlights the role of individual knowledge, context and cultural influences in shaping perceptions [225].

Analogue to machine learning embeddings, similarity metrics capture mental representations by quantifying the perceived distance between different stimuli in a psychological space. Researchers have developed various methods to quantify this distance, where two common measures are cosine similarity and Euclidean distance. According to psychological theories, such as the theory of conceptual spaces [226], individuals represent knowledge in a multi-dimensional space where similar concepts are located closer together. These metrics reflect the cognitive processes underlying categorization and perception, providing insights into how individuals mentally organize and relate different elements of their environment. By translating qualitative perceptions into quantitative measures, similarity metrics enable the modeling of complex mental representations, thereby offering a valuable tool for understanding and predicting human behavior and preferences in various contexts. While previous urban perception studies such as Place Pulse [22] have relied on ratings of predefined perceptual attributes (e.g., safety, beauty, quietness), similarity analysis differs in that it does not pre-specify categories. Instead, it uses the judgments to build an agnostic embedding space where dimensions emerge jointly from images and human responses. This design avoids constraining participants to particular concepts and allows the embedding to capture unanticipated perceptual dimensions, albeit with the trade-off that the resulting space is less directly interpretable.



### 6.2.3 Similarity in computational psychology

Similarity judgments have been extensively studied to understand how people perceive and categorize different objects based on their attributes. [175] applied the notion of similarity judgments together with machine learning embeddings with the goal of elucidating mental representations people hold about objects. To do so, they developed a triplet odd-one-out task experiment where people had to compare images of three different things. In each comparison, respondents had to choose the object that stood out from the other two. For instance, given a cat, a dog, and a coffee machine, people are expected to choose the coffee machine as it is the most dissimilar in the triplet. While

pairwise similarity ratings on a Likert scale are one of the most popular approaches, [175] argue that triplet comparisons highlight the relevant dimensions that make two things most similar. After capturing millions of triplet responses, they trained an embedding model to create the object representation.

By incorporating human similarity judgments, embedding models can provide a more complete understanding of the objects in question. Specifically, the dimensions in this embedding space capture not only visual and objective dimensions but also conceptual and subjective dimensions. This is particularly relevant for urban space embeddings, as it allows for the inclusion of nuanced human perceptions in the analysis of urban areas. By integrating these subjective dimensions, richer urban embeddings can be produced to understand how different urban environments are perceived by people, leading to more informed and people-centered urban planning and policy decisions.

## 6.3 Data

This section describes the three different types of data required for implementing our method: (1) street-level images, (2) polygonal units for covering the surface of the study area, and (3) geo-tagged population density data. In addition, we describe how the data collection process is performed for our case study in the Netherlands.



### 6.3.1 Required data

**Street-level Images**

Street-level Images (SLI) consist of panoramic photographs taken at ground level, capturing the visual and structural details of urban environments. There are different SLI providers such as Google Street View (GSV), Mapillary [227], and Apple Look Around [228], which make available these types of images around the globe. In this study, we use street-level images from GSV [229], following the image ID collection method proposed by [132]. This method involves systematically gathering geo-tagged image IDs across the study area, where each of these IDs represents one 360-degree SLI. Finally, each 360-degree image is decomposed into four 90-degree views. Figure 6.2 shows an example of this decomposition process, illustrating how a 360-degree panoramic image is divided into four separate images to provide a detailed visual coverage from different angles.

**Polygonal areas**

Polygonal areas are geometric shapes that partition a given surface into distinct and manageable spatial units. In this study, these polygons represent the minimal unit of urban space to be mapped in an embedded space. To achieve this, we utilize Uber H3 (Hexagonal Hierarchical Spatial Index), which provides a framework for spatial indexing and partitioning [230]. Given a spatial resolution, H3 divides the globe into hexagonal grids. This division ensures that each polygonal area covers a very similar surface area. While we employ H3 in our study, any other polygonal area indexing system, such as

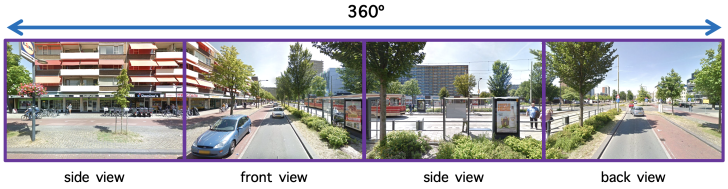



**Figure 6.2** Decomposition of one 360-degree image into four individual 90-degree images. Figure from [132]

local administrative boundaries, could also be used interchangeably. The choice of H3 is primarily made due to its flexibility in providing multiple resolutions and its global functionality, enabling consistent spatial analysis across different regions.

#### Population density

Population density refers to the number of people living within a defined spatial area. In this study, population density is used to efficiently design the urban similarity experiment by performing a pre-categorization of the urban areas based on the number of people living in them.



### 6.3.2 Data collection: the Netherlands

We collect the required data for the Netherlands by defining the spatial scope and resolution. First, we generate over one million image URLs covering the country from 2008 to 2022. These URLs are created using the Google Street View Static API [231], following the procedure described in [132]. Importantly, only image metadata (i.e., geographical coordinates and photo date) and URLs are stored. Next, we employ the H3 spatial indexing system at resolution 10 to structure the spatial data. This resolution divides the area into hexagons with approximately 60-meter sides, ensuring a fine-grained spatial representation. The country, at this resolution, is divided into about 7 million hexagons. Finally, we gather population density data from the CBS (Statistics Netherlands) *Kerncijfers* dataset [232], which contains population density information for each postal code zone in the Netherlands. These population values are then spatially joined to the corresponding H3 hexagons, enabling each spatial unit to be associated with a population density estimate. All collected data (i.e., spatial units, images and population density) is provided as input to the development of an USEM presented in the next section.

## 6.4 Method

In the following, we outline the proposed method and its implementation for constructing and training an USEM. Our approach offers a distinct solution to the *Urban Neighborhood Embedding Problem* [45] by incorporating human perceptions into its formulation. Figure

6.3 summarizes the pipeline for constructing our urban embedding in three steps: (1) image-based spatial unit definition, (2) similarity judgments collection, and (3) urban space embedding modeling. For illustration purposes, the description of the method steps is accompanied by examples from its implementation in the Netherlands. However, it can be applied in any region where the required data is available. In the following subsections, each step of the method is described in detail.

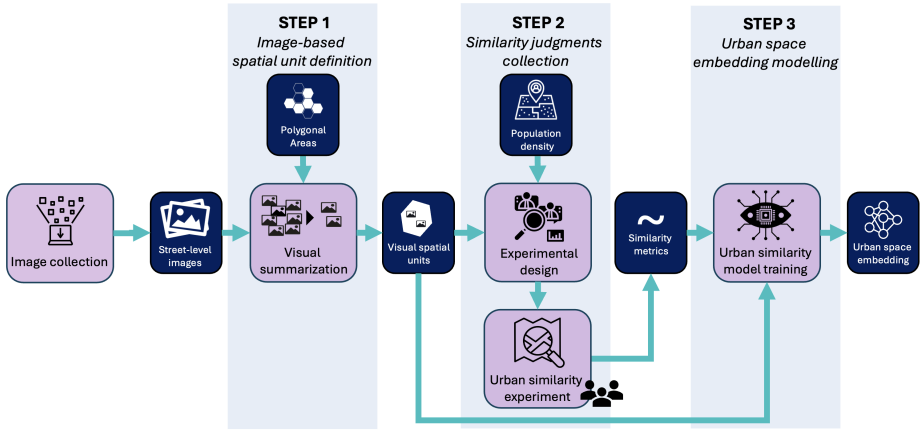



**Figure 6.3** General pipeline of the method. Purple (large) boxes are the key modeling processes in each step, and blue (small) boxes are the input/output of each sub-step. In data collection, images are retrieved. In step 1, the sampled images are spatially associated with spatial units. Then, in step 2, we collect similarity judgments through an experiment. Finally, in step 3, we use the collected similarity metrics for training the embedding model.

### 6.4.1 Step 1: image-based spatial unit definition

The first step involves visually defining spatial units by grouping and sub-selecting SLI within polygonal areas delineated by H3 at a chosen spatial resolution $sr$. H3 divides the surface of an area $\mathscr{A}$ into hexagonal units $u_i \in \mathscr{A}$, with each unit containing a unique portion of $\mathscr{A}$ and its associated SLI. In our implementation, we employ H3 polygonal areas at a resolution of $sr = 10$ and filter out hexagons containing fewer than 12 images (corresponding to three 360-degree images). This results in $780,256$ hexagons for the Netherlands.

As spatial units may contain hundreds of images, the number of images per unit is limited to a manageable subset, $k$, for facilitating practical human exploration in subsequent analyses. We develop a visual summarization algorithm for sub-selecting the images within a unit. Then, we implement this method for choosing a diverse subset of $k = 5$ images from each spatial unit, ensuring that the selected images accurately reflect the characteristics of their respective areas. This value was chosen based on practical considerations tied to our similarity judgment experiment, where participants are asked

to compare three spatial units simultaneously in a laptop screen (i.e., 15 images in total), and five images per unit provided a good balance between capturing intra-unit diversity and avoiding cognitive overload during the task. Figure 6.4 shows the procedure for obtaining $k$ subset of images from a spatial unit.

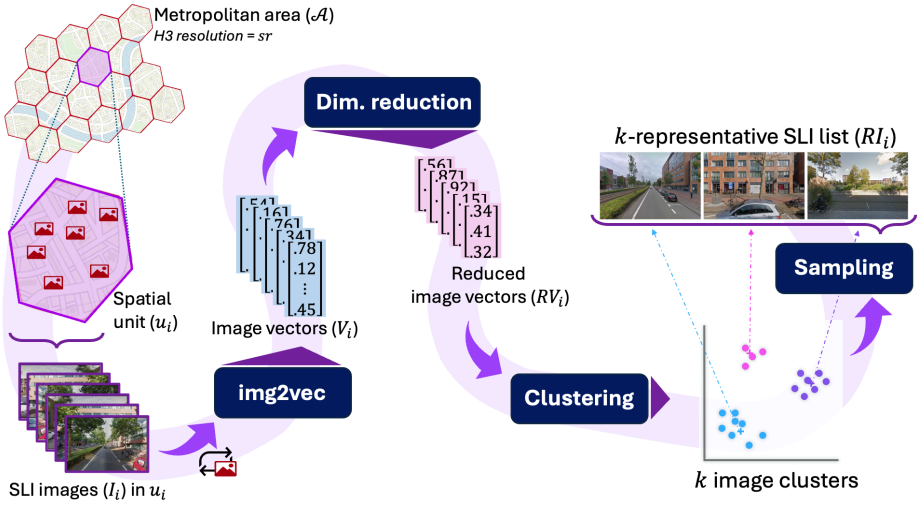



**Figure 6.4** Visual summarization algorithm used for sampling a diverse subset of $k$ images from each spatial unit. The images associated with a polygonal area are converted into vectors using any img2vec model. Then, the vectors are reduced using a dimensionality reduction technique and clustered into $k$ classes. Finally, one image per cluster is sampled.

The visual summarization algorithm comprises four sub-steps:

1. Img2vec: all images within a spatial unit are transformed into vectors (image embeddings) using any pre-trained image embedding model. For this implementation, ResNet34, pre-trained on ImageNet [25], is chosen for its simple architecture and its ability to capture complex visual features. It generates image vectors with 512 dimensions, effectively encoding detailed visual information.

2. Dimensionality reduction: high-dimensional image vectors are reduced in dimensionality to facilitate clustering. Principal Component Analysis (PCA) is used to reduce the 512-dimensional vectors to 5 dimensions while preserving around 65% of the original variance. This choice was not the result of a formal variance-retention analysis but was made pragmatically to balance computational efficiency with visual diversity and to produce manageable summaries for respondents.

3. Clustering: a clustering algorithm groups the reduced image vectors into $k$ clusters. For this study, K-means is applied with $k = 5$, ensuring that each hexagon is summarized by five visually distinct clusters.

4. Sampling: For each cluster, one representative image is selected. This can be done by choosing the image closest to the cluster centroid or selecting randomly. Here, we randomly select one image per cluster to achieve a final set of five representative images for each hexagon.

After applying these four sub-steps in all polygonal areas, each spatial unit is represented by exactly $k = 5$ images. Figure 6.5 illustrates the results of our summarization algorithm, showcasing how the original set of images within a spatial unit is distilled into five diverse and representative images.

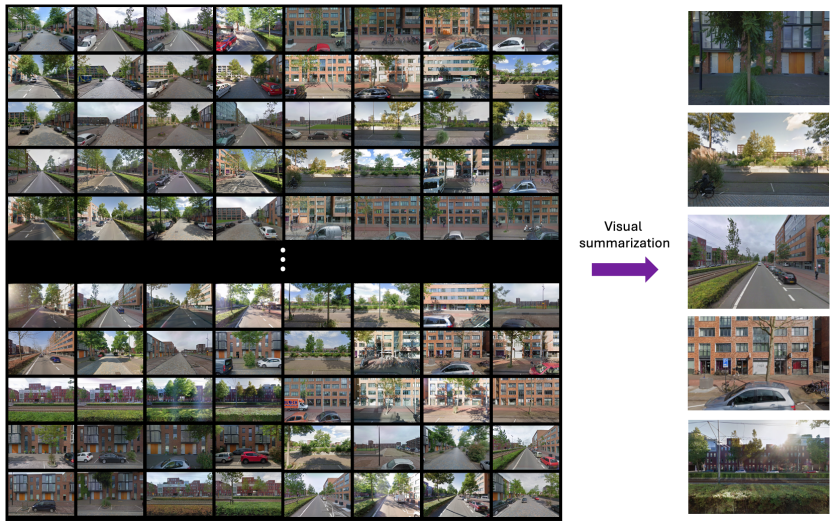



**Figure 6.5** Application of the visual summarization algorithm in a spatial unit with more than 100 images.



### 6.4.2 Step 2: similarity judgments collection

The second step involves gathering similarity judgments from people to understand how they compare different urban spaces. These judgments are collected through an urban similarity experiment using triplets of places, each represented by $k = 5$ images (defined in Step 1). In the experiment, participants are presented with several tasks. Each task displays three different places (i.e., $3 \cdot k = 15$ images), and participants are asked to select the place that is most different compared to the other two. This process involves two sub-steps: the experimental design, where we sample the places and create the triplet tasks, and executing the urban similarity experiment for collecting responses.

**Experimental design**

The human similarity experiment must be designed to maximize the informational value gained from each triplet comparison task. To achieve this, we propose focusing on two key aspects: ensuring a good representation of urban diversity and managing task difficulty. These aspects are addressed through a two-step design process: (1) place sampling, where spatial units are selected to reflect urban diversity, and (2) triplet creation, where tasks are constructed to optimize difficulty levels and their contribution to the model.

Regarding task difficulty, we assume that the more difficult a task is for a human, the more informative it is for the model. This approach is inspired by the Triplet Margin Loss introduced by [31], which is widely used for learning multidimensional embedding representations. This loss function updates the model's weights only when the triplet comparison fails to satisfy the loss margin. In other words, the more difficult the triplets are, the more information the model can gain from the training data. Conversely, if a triplet is too simple (e.g., comparing two rural areas with one highly urbanized area), the loss is satisfied, and the model learns little or nothing from the task. Additionally, difficult tasks require participants to engage in deeper analysis to identify subtle differences, which often yield more meaningful insights [233].

The experiment we carried out in the Netherlands was designed based on population density. Population density often serves as a proxy for urban development, with different density values corresponding to distinct appearances of urban areas. This allows for the pre-categorization of spatial units based on their visual appearance. Below, we describe how we use population density for sampling places and creating triplets.



1. Place sampling: we sample 10,000 hexagons (i.e., places) representative of population density throughout the country. Given the right-skewed distribution of population density, we opt for exploring the logarithm of population density, as depicted in Figure 6.6a. Next, we divide the range of log-transformed population density into five ranges, as shown in Figure 6.6a with the vertical red lines. The number and ranges are determined through visual exploration of the images within each category. As the Netherlands is a relatively small and densely populated country, there is little variation across rural and natural areas, but major diversity can be found in urbanized areas. A balanced sampling strategy ensured a representative distribution of 10,000 spatial units, with more hexagons from higher-density ranges to reflect the diversity of urban areas in population proportions. We sample $50\%$ of hexagons from range 5, $25\%$ from range 4, $15\%$ from range 3, and $5\%$ from ranges 1 and 2. This sampling strategy is designed to replicate the true population density distribution, thereby ensuring a balanced representation of population density areas, while maximizing the urban visual diversity across the Netherlands by oversampling dense areas. Finally, we construct a database consisting of 10,000 spatial units, each containing five images.

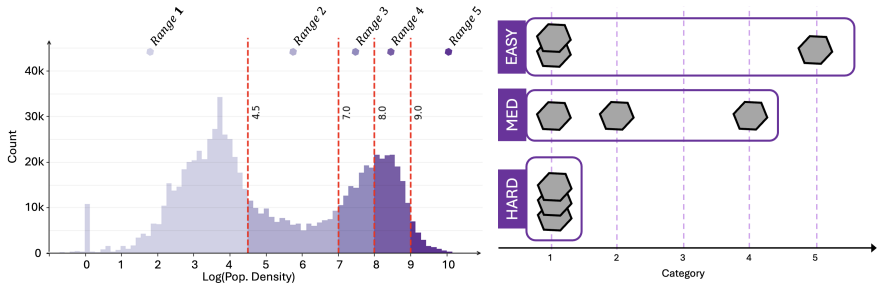



**Figure 6.6** (a) Histogram of the Log(population density) in the Netherlands. Vertical dotted lines divide the spectrum into five ranges of population density. (b) Examples of triplet tasks with different difficulties based on the population density ranges of the spatial units.

2. Triplets creation: tasks are constructed based on task difficulty, which measures the perceptual challenge of identifying the most different spatial unit within the triplet. In our application, difficulty is calculated based on the population density ranges of the three places involved using Equations 6.1, 6.2, and 6.3. The ranges come from the splits made on the population density distribution (see Figure 6.6a). The key is to measure how similar two instances are compared to how isolated the third instance is based on population density. For instance, the triplet comparison will be easier if two of the places (i.e., spatial units) come from the same range and the other one is from a different one (see the easy-marked rectangle in Figure 6.6b). On the other hand, a task is considered more difficult if the three places of the triplet are from the same population density range (see the hard-marked rectangle in Figure 6.6b).

$$minDist = \min\left(\left|R_2 - R_1\right|, \left|R_3 - R_1\right|, \left|R_3 - R_2\right|\right) \tag{6.1}$$

$$avgMaxDist = \frac{\left|R_2 - R_1\right| + \left|R_3 - R_1\right| + \left|R_3 - R_2\right| - minDist}{2} \tag{6.2}$$

$$difficulty = Rank\left(\frac{minDist + 0.1}{avgMaxDist + 0.1}\right) \tag{6.3}$$

Mathematically, the difficulty of a task is determined by Equation 6.3, where $minDist$ (Eq. 6.1) corresponds to the smallest difference in population density range indices among the three pairs of places in the triplet; and $avgMaxDist$ (Eq. 6.2) corresponds to the average of the two largest pairwise range differences in the triplet. This is mathematically represented by subtracting the smallest pairwise

distance from the total sum of distances, as it is the same as averaging the largest distances. We then compute the ratio between $minDist$ and $avgMaxDist$ and rank all unique ratio values to obtain a discrete difficulty score. In our study, we evaluated this metric across all possible triplet combinations (120 in total), which resulted in nine unique difficulty ratio values. These were then ranked and categorized into a difficulty scale from 1 (easiest) to 9 (most difficult). A small constant ($0.1$) was added to both the numerator and denominator to avoid division by zero and to differentiate edge cases (e.g., distinguishing [R1, R1, R2] from [R1, R1, R5]).

**Urban similarity experiment**

Once the triplets are created, respondents are assigned to a set of tasks with varying difficulty levels for collecting similarity judgments about urban places. Each participant is presented with a series of 15 tasks, which begins with a couple of very easy tasks (difficulty levels 1 or 2) to become familiar with the experiment. Following these, the tasks range from medium to high difficulty (difficulty levels 3 to 9). We developed a custom web platform using Python and Dash Plotly to facilitate the collection of judgments. This platform is hosted on an internal university server, ensuring data security and controlled access by the research team. The user interface is designed with a focus on user-friendliness, allowing participants to easily compare triplets of spatial units and select the odd-one-out on one screen. Figure 6.7 shows the interface of the web platform, where participants are presented with three spatial units and five images each. All responses are stored in an SQLite database, capturing the identifiers of the three places presented (place#1, place#2, place#3) along with the participant's choice of the odd-one-out.



The experiment was conducted in March 2024 and was approved by the Ethics Committee of our university. All respondents provided informed consent, ensuring their understanding of the experiment and its anonymity. We recruited participants through a panel data provider, Cint, which used stratified sampling to ensure that the sample was representative of the Dutch population in terms of age, gender and region. The panel data provider directed the sampled participants to our web platform to reply to the similarity experiment. In total, 1545 participants completed the experiment with 15 tasks each. This results in $21,810$ triplet comparisons, providing valuable insights into how people perceive different urban spaces. Plots in Figure 6.8 show the distribution of age, gender, and region of the participants. Dotted lines indicate the expected target values for a representative sample in the Netherlands. After collecting the data, we proceed with the modeling for building the urban space embedding.

### 6.4.3 Step 3: urban space embedding modeling

The final step of our method involves constructing and training an USEM. This model should be capable of mapping the $k = 5$ images from a spatial unit into a quantitative vector space based on human-perceived similarities. To do so, we train the model using a

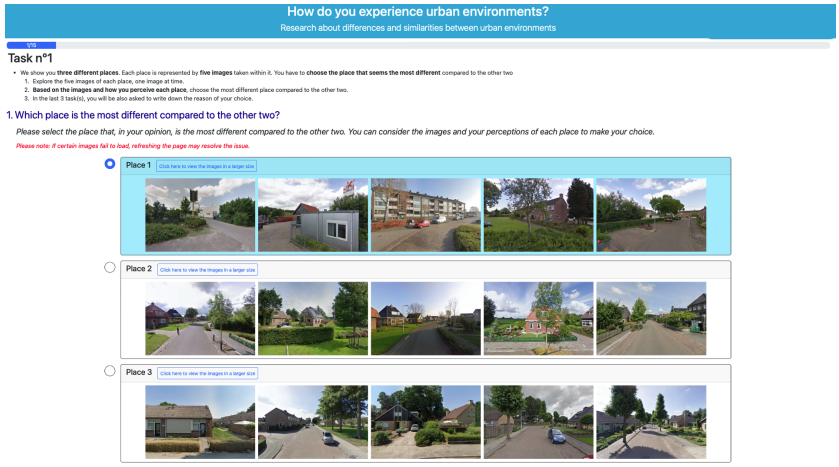



**Figure 6.7** Interface of the web platform for the urban similarity experiment. Participants are presented with three spatial units and five images for each of which, and are asked to identify the most different spatial unit.

6

triplet network architecture. This architecture is designed to process triplets of places, where each place is represented by 5 images. The triplet architecture is trained to learn the similarity metrics provided by participants through the optimization of a triplet margin loss function [31]. These similarity metrics are evaluated across all dimensions of the embedding space, allowing the model to learn rich, multi-attribute representations of urban areas. Once trained, the USEM transforms each spatial unit into a vector representation, capturing the nuanced characteristics of urban spaces as perceived by humans. The resulting embedding space allows for quantitative analysis and comparison of urban spaces based on human perceptions. In the following subsections, we detail this process divided into four main parts: defining the USEM architecture, preparing the dataset for training, defining the triplet network architecture, and training the model (based on the triplet loss function). The model architectures and training process are coded in Python using the PyTorch library and executed on an HPC using NVIDIA Tesla A100 with 80GB of memory.

### Urban space embedding architecture

The architecture of the USEM is designed to process multiple images for each spatial unit and generate a single embedding vector that captures the essence of the place. Multi-View-CNN (MV-CNN) models were introduced by [234] for processing 3D objects using a set of images, and are adopted here as each spatial unit is represented by $k = 5$ images. Figure 6.9 shows the architecture of our USEM for mapping a place with $k$ images into a unique vector space. First, we process each image individually using any img2vec model (i.e., CNN or ViT), which transforms the image into a vector representation. Specifically, in our implementation, we have tested different ResNets [25] and ConvNext

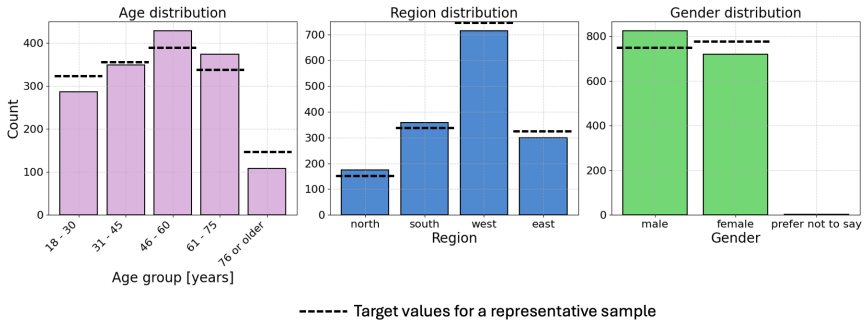



**Figure 6.8** Demographic distribution of the participants in the urban similarity experiment.

[179] versions. Once we have the $k$ vectors for the $k$ images of a place, these vectors are pooled to create a single vector representation for the respective spatial unit. This pooling can be performed using different strategies such as concatenation, average pooling, max pooling, or a combination of pooling methods to capture diverse aspects of the images. The pooled embedding is then passed through layers of a Multi-Layer Perceptron (MLP). These MLP layers encode the pooled vector into the desired final embedding vector space. The MLP layers help in learning complex transformations and relationships within the pooled data to produce a robust representation based on people's choices. This architecture ensures that each spatial unit, represented by multiple images, is effectively summarized into a single embedding vector that can be used for further analysis and modeling.



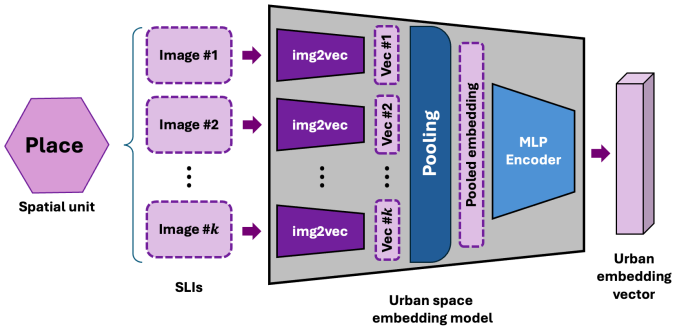



**Figure 6.9** Urban space embedding model architecture. Each spatial unit is represented by $k$ images (three in the figure), processed by an img2vec model, pooled, and encoded by MLP layers to produce the final embedding vector.

### Dataset for training

We prepare the dataset for training the USEM by combining the spatial units with the similarity judgments collected in the urban similarity experiment. Each row in the dataset contains the 15 images of the three spatial units and the odd-one-out selected by the participant. After filtering out responses that indicate random answers or speed runners, we obtain a training set with $16,654$ (90%) triplets and a testing set with $1,852$ (10%) triplets.

### Triplet network architecture for training the USEM

We employ a triplet network architecture for training the USEM presented in Figure 6.9. A triplet network consists of three identical sub-networks, each processing one of the three spatial units within a triplet. Figure 6.10 schematically presents this architecture, where each sub-network (black trapezoid) is the USEM (i.e., Figure 6.9) and transforms the $k$ images representing a spatial unit into a single embedding vector. For instance, place #1 denoted by $x_p$ has $k$ images. These images are jointly processed by the USEM to produce $f(x_p) = p$. Similarly, place #2 and #3 are processed to generate embeddings vectors $a$ and $n$, respectively. We aim at learning the weights of the USEM (black trapezoid) to satisfy the constraints imposed by the responses collected in the experiment. Specifically, we use the triplet margin loss [31], with the odd-one-out place selected by the respondent as the negative instance.



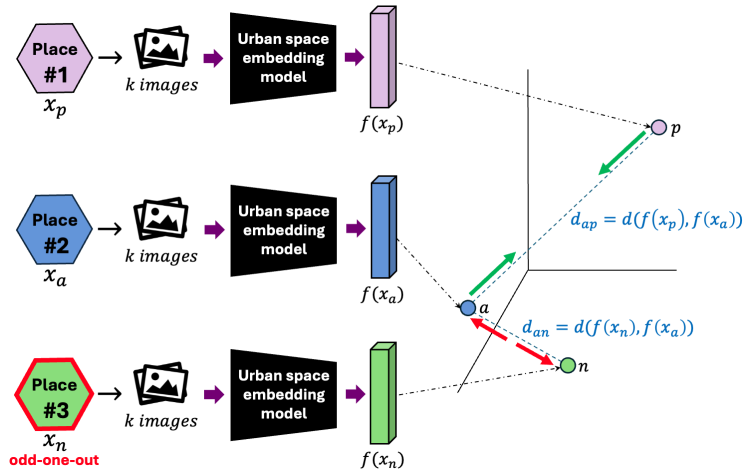



**Figure 6.10** Triplet network architecture for training the embedding model. The training is based on the experiment responses (odd-one-outs), illustrated by place #3 in the figure. Each place is represented by $k$ images. These $k$ images are mapped into an embedding space.

Figure 6.11 shows the full model we define for training the USEM in the Netherlands. First, the USEM (black trapezoid) processes the three spatial units (5 images each) independently. This produces three urban embedding vectors. Then, we add some

contrastive learning layers before computing the triplet margin loss for improving the model's performance. These layers are composed of a Multi-Layer Perceptron (MLP) called a projection head for projecting the embeddings in an even lower multidimensional space, and an L2 normalization for projecting the embeddings in the unit hyper-sphere. Similarity metrics computed by the triplet loss can suffer from the curse of dimensionality, and projecting the vectors into a lower-dimensional space can improve the model's performance [235]. The L2 normalization layer was adopted from the FaceNet architecture [31] to constrain the embedding elements to be on the hypersphere and facilitate the optimization. After the L2 normalization, the triplet margin loss is computed with the three projected embeddings and the odd-ones-out (choices) made by respondents. Then, the model parameters are updated to minimize the loss function. The contrastive learning layers are trained jointly with the embedding model to learn the similarity metrics.

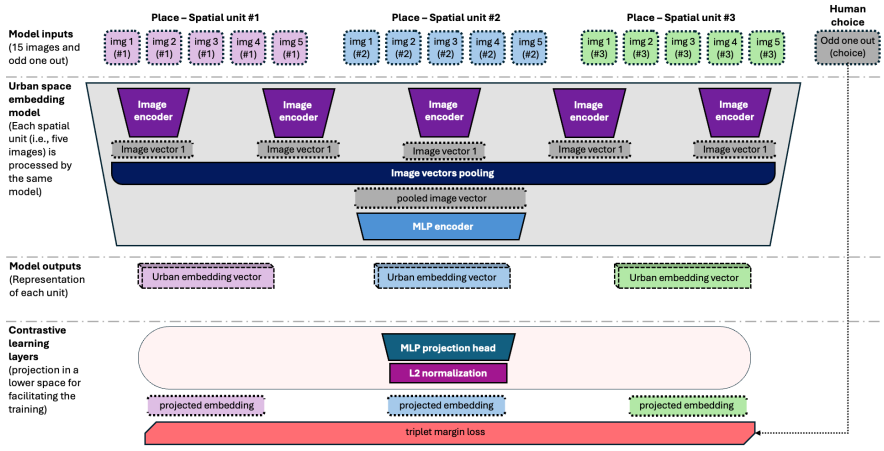



**Figure 6.11** Architecture of the triplet network used for training the urban space embedding model. The embedding model processes the five images of each spatial unit and generates an urban embedding vector. Then, the contrastive learning layers map the embeddings in a unit hypersphere for computing the margin loss and updating the parameters of the model.

### Training procedure

For training the full model, we employ the triplet margin loss function as defined by [31]. The vectors generated by the USEM must minimize the distance between embeddings of similar spatial units and maximize the distance to the embedding of the dissimilar unit, according to the triplet loss function. For instance, Figure 6.10 shows place #3 ($x_n$) as the odd-one-out, then its representation $f(x_n) = n$ in the embedding space should be located distanced from $a$ and $p$. Because in Figure 6.10 this is not the case, the USEM (black trapezoid) has to be updated to produce a new map with the distance $ap$ being shorter than $an$ (and $pn$). This loss function is designed to ensure that the embeddings of

similar spatial units are closer to each other than to the embeddings of dissimilar units in the vector space. Mathematically, this loss function aims to satisfy Equation 6.4 for each triplet $(a, p, n)$, where $a$ is the anchor, $p$ is the positive (similar to the anchor), and $n$ is the negative (dissimilar to the anchor).

$$||f(x_a) - f(x_p)||^2 + \alpha < ||f(x_a) - f(x_n)||^2 \tag{6.4}$$

In Equation 6.4, $f(x)$ represents the embedding of a spatial unit $x$, and $\alpha$ is a margin parameter that enforces a minimum separation between positive and negative pairs. The complete triplet margin loss over a batch of $N$ triplets is defined in Equation 6.5.

$$L = \sum_{i=1}^{N} \max(0, ||f(x_a^{(i)}) - f(x_p^{(i)})||^2 - ||f(x_a^{(i)}) - f(x_n^{(i)})||^2 + \alpha) \tag{6.5}$$

For training, we use the Adam optimizer to minimize the triplet margin loss function. The training process involves feeding the triplets of spatial units into the model, computing the triplet margin loss, and updating the model parameters to minimize its loss. We explore different model's architectural hyperparameters such as the MLP encoder and projection head configurations, dropout rate, triplet margin loss settings (distance metric: Euclidean or cosine; margin parameter), anchor-positive swapping during training, and weight decay to mitigate overfitting. For the training loop, we explore the batch size used to feed the triplets into the model, number of epochs, learning rates used to update different parts of the model (e.g., specific learning rate for the img2vec model. This can even be set to 0 for only training the MLP layers). By the end of the training process, the model learns to generate embedding vectors where similar spatial units are closer together and dissimilar units are farther apart in the embedding space. This trained model captures the human-perceived similarities between urban spaces and can be used for various urban analysis applications.



### 6.4.4 Model specification

The final model is selected through a hyperparameter tuning process to accurately predict choices in the odd-one-out task (from the similarity experiment). Specifically, we evaluate the model by comparing the distances between the anchor-positive and anchor-negative based on the odd-one-out selected by respondents. A prediction is considered successful if the distance between the anchor and the odd-one-out is greater than the distance between the anchor and the positive. Table 6.1 summarizes the final architecture and hyperparameters used to find out our embedding model based on the triplet architecture. Our final model produces 128-dimensional urban space embedding.

**Table 6.1** Model architecture and hyperparameters

| Aspect | Details |
|---|---|
| **Architecture** | ResNet34 $\rightarrow$ 512 (initial img. processing) |
| | Mean pooling |
| | MLP encoder: 512 $\rightarrow mlp(256, 128)$ |
| | Projection head: 128 $\rightarrow mlp(128, 64)$ |
| **Hyperparameters** | Dropout: 0.25 |
| | Distance metric: Euclidean |
| | Triplet loss Margin: 0.2 |
| | Weight decay: 0.001 |
| **Training Phase 1** | MLP layers only |
| | Learning rate: 0.001 |
| | Batch size: 256 |
| | Epochs: 7 |
| **Training Phase 2** | Full model (ResNet34 + MLP layers) |
| | Learning rate: 0.0001 |
| | Batch size: 512 |
| | Epochs: 2 |

## 6.5 Results

The main output of our approach is a model able to transform every spatial unit from the Netherlands (i.e., five SLI) into 128-dimensional embedding vectors. First, we report the performance of the final USEM model in predicting the odd-one-out place over our triplet dataset. We then apply the trained model to Rotterdam, the second-largest city in the Netherlands. This application produces the urban vectors for the city. Finally, to illustrate the information contained in the embedding space, we perform several analyses that reveal the semantics and patterns these vectors capture.



### 6.5.1 Model performance

The selected model correctly predicts 55% of the responses in the test set (and 58% in the training set). This performance is well above chance level (33%), which would correspond to random guessing. While there is no universal upper-bound accuracy for this type of perceptual task, there is a maximum achievable accuracy (noise ceiling) for any model due to the subjective nature of perceptions and internal inconsistencies of human triplet responses. Respondents may provide different answers to the same task, meaning that no single correct label always exists. Prior work by Hebart et al. [175] estimated a noise ceiling of approximately 67% in a comparable triplet-based setting, though their study focused on object images rather than urban environments and involved fewer visual stimuli per task (three images versus our fifteen). Despite these differences, the concept of a noise ceiling remains a useful point of reference: even a perfect model cannot achieve 100% accuracy in the presence of perceptual ambiguity. In our case, the greater complexity and scale of the visual stimuli likely introduce even more noise, suggesting that the theoretical ceiling may be lower. While we do not treat Hebart's estimate as a definitive benchmark, it provides a conceptual frame to interpret our results.

Our model's performance can thus be understood as capturing a substantial portion of the explainable variance in human perceptual judgments, approximately 82% of the estimated maximum accuracy.

### 6.5.2 Urban Space Embedding Model results

We apply the selected model to spatial units from the city of Rotterdam. This is to showcase the information that our embedding can capture. Rotterdam is the second largest city in the Netherlands with approximately 650,000 inhabitants and covering $132km^2$. It offers a diverse urban landscape, including high-density areas, commercial and industrial zones, as well as green spaces. This makes it an ideal case study for our model. We consider 7,332 hexagonal spatial units in Rotterdam's metropolitan area (excluding the port area), processing a total of 36,660 images. Because only $4.7\%$ of the training images came from Rotterdam and just one triplet compared two locations within the city, this application also tests how well the model generalizes to largely unseen urban visual data. Accordingly, the model produces $7,332$ 128-dimensional embeddings for each hexagon in the city. In the following sub-sections, we discuss the information captured from different angles.

#### Spatial patterns based on embedding features



We explore spatial patterns based on the embedding features to understand the distribution of urban spaces across Rotterdam. To do so, we use Principal Component Analysis (PCA) to reduce the 128-dimensional vectors to three dimensions. This allows us to represent each spatial unit with one color using RGB channels [46]. As each PCA dimension can be normalized to a range of 0-255, each PCA value serves as an RGB channel. Using this approach, we can visualize the urban space embedding in a colored map, where similar colors represent similar embeddings. Figure 6.12 shows the resulting map, with each hexagon colored based on its PCA3 embedding values.

The map reveals distinct urban zones across the city of Rotterdam, each characterized with different colors and tonalities. For instance, two main highways (i.e., highways A16 and A20), intersecting at F6, stand out prominently in a green-yellowish tone. Adjacent to this intersection, in pink, there is a large park (*Kralingse Bos*), mirroring the coloration of the northeast area of Rotterdam, which is known for its high vegetation density. The main city center, in D4, is distinguished by a lighter shade, contrasting with its surroundings. This lighter shade is also in the sub- centers located in DE2 (*Zuidplein*), FG6 (*Rotterdam Alexander*), and B5 (*Overschie*), suggesting a similarity in urban characteristics to those of the main city center. In contrast, the city's periphery, beyond the highways, takes a darker tone compared to the inner city area. This delineates the transition from the central urban fabric to the outer suburban areas. This color-coded visualization effectively captures Rotterdam's diverse urban landscape, from its center to its green spaces and sub-centers.

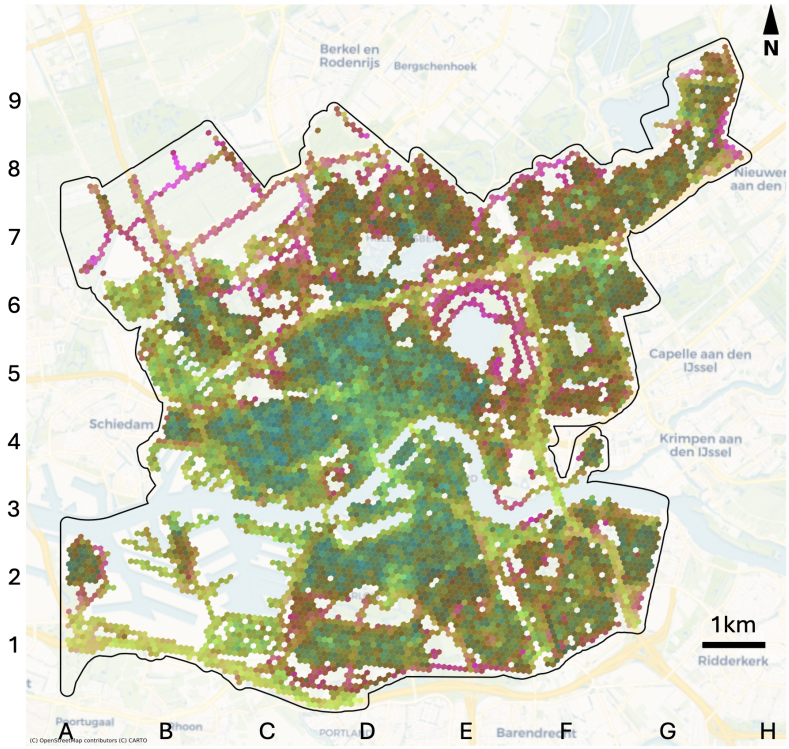



**Figure 6.12** Embedding values projected in three dimensions using PCA. Each hexagon's color is produced by normalizing the PCA values between 0 and 255 and using them as RGB channels. Similar colors indicate similar embeddings in the PCA space, and therefore similar urban spaces.

### Similarity relationships across urban areas

We quantitatively explore the similarities between different urban spaces to delve deeper into the information captured by the urban space embedding. Similarity can be measured by computing the Euclidean distance between the vectors of two urban spaces. Specifically, we measure the distance between one reference hexagon and all other hexagons in Rotterdam. Then we plot the results as a heatmap. Figure 6.13 presents four heatmaps, each illustrating similarities between Rotterdam's urban spaces with a different reference hexagon. The similarity values are derived from the Euclidean distance between the reference hexagon and all other hexagons in Rotterdam and then normalized between 0 and 1. In these heatmaps, a value of 0 (shown in red) indicates low similarity, while a value of 1 (shown in green) indicates high similarity.

The heatmaps reveal results which are consistent with the spatial patterns discovered in Figure 6.12. For instance, the city center hexagon (Figure 6.13a) shows high similarity with hexagons in the immediate vicinity, such as the adjacent hexagons in the city

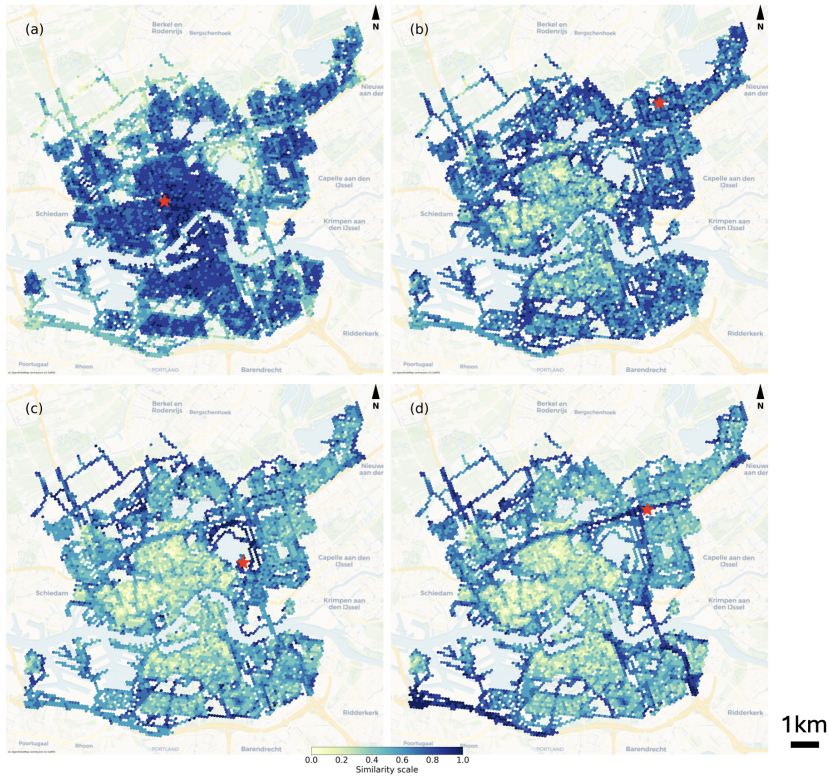



**Figure 6.13** Heatmaps showing the similarities between Rotterdam's urban spaces using four reference hexagons. Each map represents the similarity of one hexagon (marked by the white star) with all others in Rotterdam. (a) hexagon in the city center, (b) hexagon in the periphery, (c) hexagon in a park, and (d) hexagon in a highway.

center and some clusters located in the suburban areas (which are also light-colored in Figure 6.12). The periphery hexagon (Figure 6.13b) has a low similarity with the city center, but a higher similarity with other periphery hexagons. The park hexagon (Figure 6.13c) is most similar to hexagons in the northeast area, where green spaces are prevalent. Finally, the highway hexagon (Figure 6.13d) shows high similarity with hexagons along the highways and the northeast area. These heatmaps reveal clear patterns that align with the spatial distribution explored in the previous subsection and additionally provide a quantitative measure of similarity between different urban spaces. For example, comparing the suburban hexagon (Figure 6.13b and the city center hexagon (Figure 6.13a) with a park area, the suburban hexagon is around 30% more similar. This may indicate that the suburban hexagon has more vegetation compared to the city center hexagon.

### Zonification of spatial units

We also delineate different zones based on the proximity of units and embedding features. While the map in Figure 6.12 visually reveals the existence of different zones through varying colors, it remains difficult to accurately identify the number of zones and their precise boundaries. Clustering the urban vectors facilitates the identification of these zones and the understanding of their characteristics more clearly. We employ sequentially two clustering algorithms, agglomerative and k-means, to identify distinct zones within Rotterdam. Initially, we apply agglomerative clustering, taking into account both the similarity across urban vectors and hexagons' adjacency. This allows us to identify groups of adjacent hexagons with similar embedding characteristics. We set this algorithm to produce a very high number of clusters (i.e., 200 clusters), each maintaining internal similarity. Next, we compute the average vector for each cluster to derive a single representative vector for each cluster. Finally, we apply k-means clustering on these average vectors to determine five distinct zones within Rotterdam. We selected five clusters for illustration purposes, as increasing the number of clusters would identify more regions but result in less pronounced visual distinctions (i.e., it requires more images to show the differences). In this way, we first split the city into many semantically similar and geographically adjacent zones, and then we cluster these zones to group them without the geographical constraints. This allows us to determine the similarities across the zones using a lower number of clusters (i.e., five clusters). The map in Figure 6.14 shows the results of the clustering analysis, with the urban vectors color-coded according to the identified zones. Additionally, the photos on the right side are randomly sampled are from hexagons near the centroid of each zone.



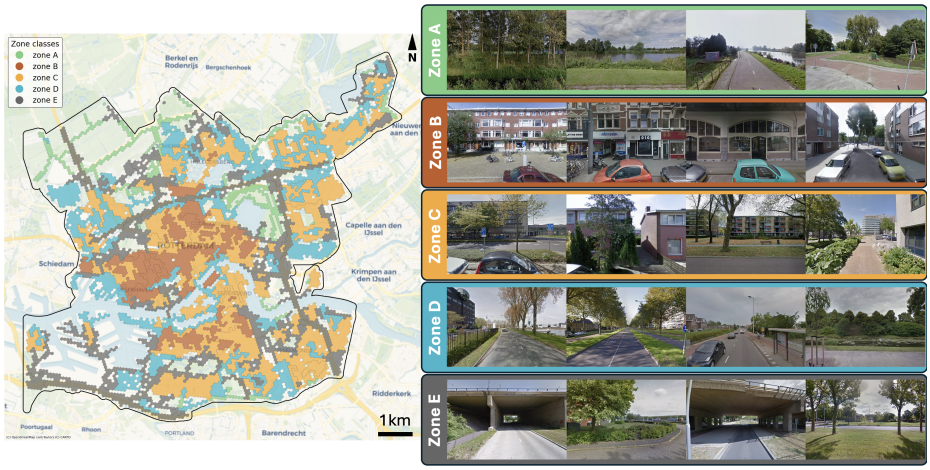



**Figure 6.14** Zonification of Rotterdam based on the urban space embedding features. The urban vectors are color-coded according to the identified zones, and random images are sampled per zone.

Figure 6.14 presents the results of the zonification of Rotterdam based on urban space embedding features and the clustering techniques. This process divides the city into five distinct zones, each represented by a unique color. The inner city is primarily covered by zones B and C, colored brown and orange, respectively. Zone B, predominantly in the inner city, features a dense urban fabric with commercial areas, orange-brick buildings, and limited vegetation. In contrast, zone C includes more open spaces with grasslands and trees, indicating a slightly less dense environment. Parks and natural areas are effectively clustered as well. Consistent with Figures 6.12 and 6.13c, areas with abundant vegetation and natural water bodies are grouped together in Zone A, shown in green. The periphery of the city is mainly divided into zones C and D, represented by orange and blue. While Zone D shares similarities with Zone C, it reveals an even greater presence of trees and individual houses, indicating a suburban residential character. Finally, the highways and main arterial roads are clustered together in Zone E, depicted in gray, which aligns with the geographical distribution of highways in Rotterdam. Also, zone E includes some industrial and port areas, as can be seen in the south-east of Rotterdam.

**Exploration of the multidimensional space**

We also explore the structure, shape, and semantics of the multidimensional urban space embeddings. To this end, we apply t-SNE for reducing the 128-dimensional vectors into two dimensions. This technique is well-known for preserving the local structure of the data, which allows for a visual inspection of the data. The left side of Figure 6.15 shows the t-SNE projection of the urban space embeddings in Rotterdam. This 2D visualization is colored according to the zones identified in Figure 6.14. Additionally, to examine the transitions within the embedding space, we select two random hexagons located at the boundaries of the vector cloud and connect them with a line. We then identify the closest hexagon to the line at five equally spaced points along this path and display the corresponding images. Figure 6.15 presents the t-SNE projection of the urban embeddings on the left and the five sampled hexagons along the selected path on the right.

The t-SNE visualization in Figure 6.15 effectively groups the zones described in Figure 6.14, validating the similarities identified by the clustering analysis, as hexagons from the same zone appear adjacent to each other in the t-SNE projection. The shape of the entire space provides insights into the structural organization of urban spaces in Rotterdam. Notably, the point cloud displays three main clusters on the right side: the brown (zone B), more city-centric, gray (zone E), as highways, and green (zone A), as high-vegetation areas. These clusters represent the most distinct urban zones, with the zone C cluster (orange) acting as a transitional area among them. Zone C emerges as the most prominent and diverse zone in Rotterdam, as it relates to all other types of residential areas.

The line connecting the selected boundary hexagons illustrates the gradient between different urban spaces. The five sampled hexagons along this line reveal visual transitions from an urbanized area (hexagon 1) to a green space (hexagon 5). Significant transitions

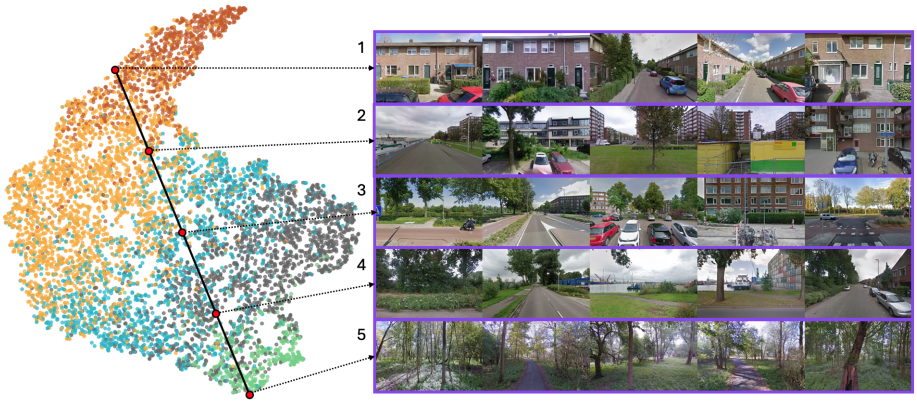



**Figure 6.15** Exploration of the multidimensional urban space embeddings in Rotterdam. The left side shows the t-SNE projection of the embeddings colored by the identified zones. The right side displays images from hexagons along a path connecting two random hexagons in the embedding space.

include an increase in vegetation and openness between buildings. More subtle changes are observed in the transition from individual houses to larger buildings, then to an industrial zone, and finally to a park devoid of buildings.



### Added value of human perceptions in urban embeddings

We conduct an analysis to assess the added value of incorporating human perceptions into the model. This is achieved by using a ResNet34 model pre-trained on ImageNet, without fine-tuning on human responses, to generate urban space embeddings. This baseline model processes images through ResNet34 followed by max pooling to produce a vector for each hexagon. We then compare the accuracy of predictions from this ImageNet-based model with those obtained from the model trained on human responses. The ImageNet model achieves an accuracy of 48%, while the model trained with human responses achieves an accuracy of 55%. This 7 percentage point difference indicates that incorporating human responses enhances the model's ability to reflect how humans perceive urban spaces by approximately 14.5%. This is not an indication of one model being inherently superior to the other, but rather a demonstration that embedding human perceptual judgments introduces an additional alignment with lived human experiences that a vision-only model cannot capture.

We apply the zonification approach used earlier to the embeddings generated by the ImageNet-based model. Figure 6.16 displays the clusters identified using the ImageNet model. We try to replicate the colors used in Figure 6.14 to ease their comparison. We also sample images from hexagons in the inner city, specifically from the delineated zone

G on the map. This area encompasses zones B and C from Figure 6.14, to which this model assigns a unique cluster. The image samples include zones B and C (see Figure 6.14).

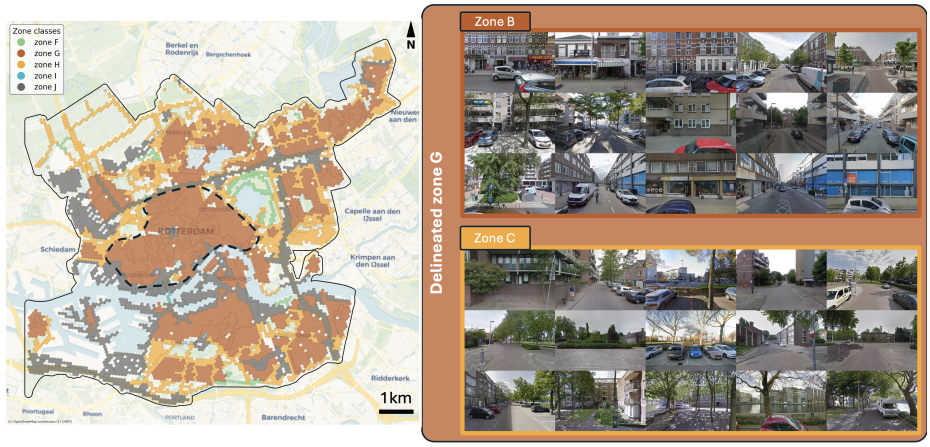



**Figure 6.16** Comparison of model's predictions and the participants' choices in the urban similarity experiment. The map shows the accuracy of model's predictions in relation to participants' choices for each hexagon in Rotterdam.



The ImageNet-based model provides a good general classification of Rotterdam's urban landscape, identifying major features such as the city center, highways, green spaces, and suburban areas. However, it clearly lacks the subtleties and nuances needed for detailed differentiation within these categories. Figure 6.16 demonstrates that the ImageNet-based model can effectively identify four primary clusters: residential city-centric (zone G in brown), residential suburban (zone H in yellow), parks (zone F in green), and highways (zone J in gray). A fifth cluster is largely classified as noise (zone I in blue). This limitation highlights our model's ability to differentiate between the nuanced characteristics of the urban environment.

In contrast, the model trained with human responses reveals nuanced groupings within these main urban categories. Specifically, images from the delineated zone G, which encompasses zones B and C from Figure 6.14, highlight a clearer division within this delineated zone. Zone B, characterized by a city-centric layout with commercial areas and narrower streets, contrasts with Zone C, which features a more open, green area with greater vegetation and lower building density. This indicates that while the ImageNet-based model is sufficient for general urban space classification, a model with human perceptions offers a considerably more refined depiction of how urban space is perceived.

## 6.6 Discussion

We have proposed a novel method for enhancing urban space embeddings by integrating human perceptions into their formulation. In addition, we developed an algorithm for the visual summarization of spatial units, which systematically defines urban areas through images. We also designed an experiment and a web platform to collect human similarity judgments about urban spaces, and we created a model that trains urban space embeddings based on these perceived similarity metrics.

Our research adds to the growing body of work focused on understanding and modeling urban environments through human perception. For instance, prior studies such as Urban Mosaic by [236] have demonstrated the potential of large-scale visual data in exploring and differentiating streetscapes, while [212] emphasized the importance of incorporating safety perceptions into urban analysis. Building on these insights, our approach further advances urban space modeling, offering a more nuanced and human-centered representation of urban environments.

By collecting human similarity judgments and integrating them during the training phase, we effectively captured how different urban spaces are perceived by people. The experiment involved 1,545 participants, generating 21,810 triplet comparisons of urban places. We trained our USEM with this data, achieving a prediction accuracy of 55%, which represents approximately 82% of the estimated best possible accuracy, considering the inherent noise in perception-based tasks [175].



Applying this model to Rotterdam, we produced 7,332 spatial unit vectors of 128 dimensions. The model successfully captured patterns across various urban environments within the city, distinguishing residential areas, commercial zones, green spaces, and highways. Our results in Rotterdam demonstrate the added value of integrating human perceptions into urban embeddings. The model trained with human responses outperformed an ImageNet-based model (without human input) in considering people's perceptions for clustering urban areas. This highlights the relevance of incorporating human insights to enhance the model's ability to disentangle urban nuances more accurately.

However, we acknowledge several limitations. Images shown on a screen cannot fully capture the sensory experience, ambiance, and nuanced characteristics of real-world places, such as sounds, smells, or a sense of security in those spaces. This discrepancy highlights the challenge of translating complex urban environments into visual representations for analysis. In addition, the images used in our analysis were sourced from Google Street View, primarily from areas accessible by car, which may not fully represent pedestrian experiences or other urban characteristics. Furthermore, we did not control for weather conditions in the images, which could introduce weather-related biases into the experiment. We also did not control for prior knowledge participants may have about the locations and how this could influence their responses. Familiarity with the places might lead them to consider contextual factors beyond the visual content of the

images, potentially introducing a bias in the results. Another limitation is that the triplets were primarily created based on population density, which could potentially bias the learning process by overlooking other characteristics related to population density.

Additionally, the design of the triplets in our experiment posed significant challenges. The number of images used for representing each spatial unit was directly connected to the design of the triplets in our experiment and it may not always capture the full heterogeneity of urban areas. The similarity experiment should have the right balance between difficulty and informational value. We finally considered five images per spatial unit to ensure adequate representation, but this number of images could make some tasks too difficult for people. Also, the training phase required informative triplet comparisons to effectively learn differences among urban areas. To support this, we developed a heuristic difficulty index based on population density ranges, which served as a proxy for estimating the cognitive effort required to respond to each task. This offered a practical solution to improve the efficiency of data collection by balancing information retrieval per task. Although this index is not a validated measure of perceptual complexity, other psychological techniques could be explored. Related to the survey data collection, we did not include systematic measures for reliability tests, therefore, future implementations could incorporate tester triplets with expected majority answers or repeated triplets to more formally assess response consistency. Additionally, training the triplet network model was particularly difficult due to its high degree of flexibility in satisfying triplet constraints [31], which often led to a collapsed model where the loss converged to the margin value, resulting in sparse embedding vectors [31]. To address this, we employed larger batch sizes [235] and incorporated L2 projection before computing the loss [31], which stabilized the training and reduced the likelihood of model collapse. Finally, on comparing USEM to a baseline ResNet model to isolate the effect of human feedback. However, a broader evaluation against existing state-of-the-art urban embedding models and through downstream tasks could offer further insights into its practical value and generalizability.



These limitations also offer different opportunities for future research. The proposed method entails a trade-off: the urban embedding is agnostic to predefined perceptual categories, which allows it to capture unanticipated dimensions but makes it more difficult to interpret directly. Post hoc techniques, such as analyzing participants' stated reasons for their choices, offer one possible avenue for probing which dimensions are being represented, and we see this as a promising direction for future work. Also, exploring how different people's backgrounds and socio-demographic factors might influence the embedding spaces could reveal how diverse populations perceive urban environments differently, leading to more inclusive and targeted urban models. Additionally, it would be valuable to experiment with other multi-modalities to assess whether more types of data can reduce the need for human input by capturing intrinsic differences between urban spaces. Incorporating new technologies into the similarity experiment could also make it more realistic and insightful. For example, studies such as [237] and [238] have explored the potential of VR-based platforms for capturing perceptions of urban spaces, while [239] utilized eye-tracking technology to identify areas that draw human attention

in public spaces. Another promising avenue is extending the similarity experiment to capture perceptions of change over time. As demonstrated by [240], street-level images can be leveraged to model the evolution of urban environments. Applying a similar approach within the context of our similarity experiment could provide insights into how human perceptions shift in response to urban transformations, such as gentrification, infrastructure upgrades, or environmental changes. Finally, the embedding results of this work could serve as input for other applications such as analyze boundaries of neighborhoods semantically, or quantify changes over time in term of perceptions or physical things. This could further enhance the ability of urban embeddings to reflect not only static attributes but also the dynamic nature of cities and their evolving social landscapes.

Incorporating human perceptions into urban modeling enables a more human-centered analysis of cities, enhancing both the nuance and relevance of urban space embeddings. This study demonstrates the potential of integrating computer vision with human perceptual data to develop models that better capture how people experience urban environments. While pre-trained vision models offer useful insights for general analyses, adding human input allows for a deeper understanding of spatial patterns and the impacts of urban interventions. For instance, our similarity experiment could be applied to evaluate new projects based on the perceptions of those directly affected, thereby supporting decision-making processes that prioritize people's needs and experiences. Beyond this, perceptual embeddings such as USEM could support a range of applied urban analytics, including identifying perceptual divides between neighborhoods, measuring the perceptual impact of urban renewal, or refining spatial boundaries based on how people mentally group areas. This approach opens new possibilities for urban planning, policy-making, and research, highlighting the growing importance of perception-driven analytics in urban design.



## 6.7 Conclusions

This study introduced a new method for representing urban spaces as numerical embeddings derived from SLI, explicitly integrating human perceptions into the modeling process. The method captured multi-attribute characteristics of urban spaces, including both perceptual and visual information, and used them as input for training a triplet network model. In doing so, it addressed a key gap in the existing literature: the lack of human-centered information in urban representation modeling.

The resulting 128-dimensional embeddings notably distinguished between different urban settings without relying on predefined labels. This offers a richer and more precise spatial categorization. It is important to note that the model does not yet identify specific perceptual qualities (e.g., safety or vibrancy) for each spatial unit; instead, it captures differences that emerge from aggregated human similarity judgments. Additionally, this model enables a variety of downstream tasks, such as zoning analysis, exposure assessments, and change detection, while considering human perception.

Nevertheless, this work has some limitations and opportunities for future work. The modeling approach did not consider other sensory input beyond visual; control for participants' prior knowledge about certain locations; or explored generalization of the model across other countries. Related to the data, the use of SLI could introduce biases by oversampling car-accessible areas and reflecting uncontrolled environmental conditions like weather and lighting. Future work could address these limitations by incorporating multi-sensory data, applying the methods across diverse global contexts, and refining experiment design to better manage cognitive load while preserving perceptual information. Overall, this research contributes a new perspective to urban analytics—one that centers human experience as a core element in the spatial representation of cities.

## Data and codes availability

The data, codes, and instructions that support the findings of this study are available with the identifier(s) at the private link: `https://github.com/FGarridoV/From-pixels-to-perceptions`

# Chapter 7

# Computer vision-enriched discrete choice models


## Abstract

The final study integrates components, visual perception and behavioral modeling by proposing Computer Vision–Enriched Discrete Choice Models (CV–DCMs). These models extend traditional Random Utility Maximization frameworks to incorporate visual information extracted from street-level imagery. In a residential location choice experiment, respondents evaluated alternatives described by both numerical attributes (e.g., cost, commute time) and images of street-level conditions. Image embeddings derived from a visual feature extractor were combined with choice data to estimate how visual characteristics influence preferences. The results demonstrate that perceptual features—such as greenery, enclosure, or visual order—significantly shape residential desirability, complementing traditional quantitative factors. Methodologically, the CV–DCM represents a new integration of behavioral choice theory and computer vision, enabling utility-based models to learn from image data directly. This approach couples pixels with people and places, and positioning computer vision as a foundation for perception-aware behavioral modeling in urban analytics.


**Note: The author contributed to the data collection process, model conceptualization, model implementation and training. The author did not contribute to the discrete choice modeling part of this study.**


This chapter is based on the journal article: van Cranenburgh, S., & Garrido-Valenzuela, F. (2025). *Computer vision-enriched discrete choice models, with an application to residential location choice.*Transportation Research Part A: Policy and Practice, 192, 104300.

## 7.1 Introduction

Discrete Choice Models (DCMs) are widely used in transportation (and beyond) to describe how individual choices result from preferences over attributes and available alternatives in multi-attribute decision-making. When DCMs were incepted in the 1970s, they were used to explain and predict mode and destination shares [241, 242]. Nowadays, DCMs are applied to a wide variety of choice situations, including residential location choice, route choice, vehicle choice, airport choice, time of day choice and many more [243–249]. DCMs are built on the notion that attributes have numeric values or can be converted into numeric values, e.g. in the case of a categorical level. In other words, the attributes that jointly make an alternative only involve numbers.

Visual imagery is crucial to many multi-attribute decision situations, in and beyond transportation. For example, visual information is indispensable to residential location choices. In today's digital age, it is hard to imagine searching for a house on a real estate website without access to images. Other examples of such decision situations in transportation include vehicle choices, tourist destination choices, transport infrastructure design choices and choices related to safety, such as where to cross a street on foot and whether a route is safe enough to cycle. The widespread use of visual imagery, e.g. on websites like *Zillow.com* and in Stated Choice (SC) experiments, can be attributed to the fact that it is easier for people to perceive and process information presented through images than information presented in text or numbers [172]. In addition, visual imagery provides valuable details about the alternative, such as scale, texture, or quality, that are difficult to convey through textual descriptions or numbers [250]. For instance, in a residential location choice context, visual characteristics of the (built) environment (henceforth referred to as street-level conditions) such as "safeness", "openness", "continuity", and "common orientation" cannot be easily expressed in numbers but can effectively be communicated by images. The COVID-19 pandemic brought the importance of street-level conditions to the forefront, with millions of white-collar workers relocating to suburban areas with better street-level conditions during the pandemic-induced remote work shift [251, 252]. Therefore, to accurately represent choice behavior in multi-attribute situations that involve visual imagery, it is necessary to have choice models capable of working with image data.



However, present-day DCMs cannot handle image data directly and, therefore, cannot incorporate information from images into their representations of choice behavior. The inability to handle image data in DCMs creates a stark contrast between the behavior it seeks to model, where images are widely used, and what DCMs can do. Even when researchers deliberately use images in SC experiments to visualize information that is challenging to convey in numbers, the information embedded in the images is scantly accounted for ([253]; see Hevia-Koch and Ladenburg [254] for a thorough discussion). DCMs' inability to handle image data leads to incomplete and potentially misleading outcomes.

As a solution, this study proposes "Computer Vision-enriched Discrete Choice Models" (henceforth abbreviated as CV-DCMs). These models can handle choice tasks involving both numeric attributes and an image. CV-DCMs are grounded in Random Utility Maximization (RUM) principles [255, 256]. Therefore, CV-DCMs maintain the solid behavioral foundation of traditional DCMs while expanding their application to include image data. We demonstrate the effectiveness of the proposed CV-DCMs by shedding light on the importance of street-level conditions to residential location choice behavior relative to travel-related factors, such as travel time and travel cost. To do so, we have developed and administered a novel stated choice experiment involving trade-offs between commute travel time, monthly housing cost (both numeric attributes) and street-level conditions (presented using images).

The main contribution of this paper is methodological. It contributes to the growing body of literature in the travel behavior field that seeks to integrate machine learning and DCMs (e.g. [50, 51, 257–262]). More specifically, our study can best be positioned in two streams of the literature. The first stream of literature concerns studies that map human perceptions of the urban environment using a combination of street view images and machine learning [22, 47, 137, 213, 263, 264], and see Ito et al [265] for a review). In this stream of literature, models are trained on survey data in which respondents are typically presented with two street view images and asked to indicate which image looks safer/more vibrant/livelier/etc. After training, these models are commonly used to generate spatial maps showing where the urban environment is perceived as safe/vibrant/lively/etc. Our study also uses street view images and machine learning but deviates from this literature because it concerns preferences. Although perceptions and preferences are closely related concepts, they are not the same. Preferences are grounded in the theory of choice behavior [266–268] and govern what people choose and how they make trade-offs. In contrast, perceptions are subjective interpretations of sensory stimuli, which may influence but do not necessarily determine individuals' choices [269].



The second stream of related literature concerns studies seeking to understand choice behavior (and thus preferences) in the presence of visual stimuli (i.e. images) by first encoding the information from images into tabular form and then estimating traditional choice models. Encoding information from images can be done manually by the researcher (see e.g. [270, 271]) as well as algorithmically by computer vision algorithms. Manual encoding is labor-intensive and imposes a strong limitation on the number of images that can be utilized. Noteworthily, Patterson et al. [272] circumvent this challenge using artificially created images that reflect specific attribute levels, such as dwelling type and space between buildings. Studies taking the algorithmic encoding approach typically use object detection and semantic segmentation models to extract information from images (e.g. [50, 51]). Directly encoding information from images into tabular form offers a significant advantage; the modeling results (i.e. the preference parameters) are directly interpretable. However, this approach critically relies on prior knowledge of the factors influencing (choice) behavior and the accuracy of the information extraction (either by human annotators or algorithmic object detection and segmentation models). The model proposed in this study does not rely on prior knowledge of the factors

influencing the choice behavior or the accuracy of object detection models. Also, it preserves the interpretability of the model's parameters associated with the numeric attributes. However, because our model is trained end-to-end and its encoding is 'hidden', insights regarding preferences over images cannot be derived from scrutinizing the model's parameters.

Finally, this research substantively contributes to the residential location choice behavior literature. Specifically, it shows the importance of street-level conditions in residential location choices relative to commute time and housing cost. Additionally, it sheds light on the heterogeneity in preferences over street-level conditions. These substantive insights can be valuable for informing urban planning and housing policies.

The remaining part of this paper is organized as follows. Section 7.2 describes the proposed CV-DCMs. Section 7.3 discusses the stated choice data collection effort and reports the sample statistics, descriptive results and details on the training of the model. Section 7.4 contains the main results. Section 7.4.1 presents the results from the CV-DCMs and compares model fit and parameter estimates with those of traditional discrete choice models, which do not account for images. Section 7.4.2 shows what the CV-DCM has learned about what decision-makers find relevant for their residential location choices. It provides face validity to the modeling results. Section 7.4.3 demonstrates the merits of the CV-DCM by showing how CV-DCMs can be used to deepen understanding of residential location preferences. Finally, Section 7.5 draws conclusions and discusses limitations and directions for future research.

## 7.2 Method



This section presents the methodology. Section 7.2.1 introduces relevant models and concepts from computer vision. Section 7.2.2 proposes the modeling framework. Section 7.2.3 briefly discusses implementation details and training.

### 7.2.1 Preliminary: computer vision models and concepts

Computer Vision (CV) is concerned with extracting meaningful information from images, videos, and other forms of visual data. CV models typically detect scenes and objects in images [273]. Nowadays, CV models are applied in a wide range of applications and numerous fields. In transport, CV models are essential for future autonomous vehicles to perceive and understand their environment; in healthcare, CV models are used in medical imaging to aid in diagnosing diseases and abnormalities; and, in retail, CV models are used to track customer movement in stores. As CV models grow and become more powerful, they can perform increasingly sophisticated visual tasks [274]. The largest CV models currently in use contain over 1 billion weights [275].

The building blocks of images are pixels. A pixel represents a single point in an image and contains information about its color and brightness. Each pixel has a spatial location ($hxw$) and a color value. Most color images nowadays use three color channels: Red (R), Green (G), Blue (B), and 8 bits per color channel (implying three 0-225 values),

with which it is possible to create a wide range of colors and shades. Mathematically, images are usually represented as 3D tensors, which are multi-dimensional arrays of numerical values. Tensors enable easy processing and manipulation of images using various mathematical operations and algorithms, especially in combination with GPUs. The three dimensions of an image tensor typically correspond to the image's width, height, and color channels. Thus, an RGB color image with a resolution of 900 x 600 pixels can be represented as a 3D tensor with a shape of (900, 600, 3), where the first two dimensions correspond to the height and width of the image and the third dimension corresponds to the color channels. An image tensor of a 900 x 600 RGB color image contains 1.6m data points.

CV models typically have two main components: a feature extractor and a classifier, see Figure 7.1. The feature extractor is generally a deep neural network that is trained to extract relevant features from images. The output of the feature extractor is the feature map, which is a lower-dimensional vector representation of the image and captures its salient features. In other words, the feature map contains (most of) the information of the image but is more compact in form. Usually, a feature map (a.k.a. embedding) is a flat array of floating points. Nowadays, a so-called transformer architecture is the mainstay choice as a feature extractor in the CV field [2]. In contrast to traditional convolution-based architectures, so-called Vision Transformers (ViT) rely on self-attention and multi-head attention mechanisms to learn spatial relationships between different parts of the image. In a ViT architecture, the input image is divided into a grid of non-overlapping patches, which are linearly embedded to produce a sequence of feature maps. These feature maps are then processed by a series of transformer encoder layers to learn spatial relationships between the different parts of the image. The classifier is a separate component which is trained to classify the input image based on the feature map. Typically, the classifier is a Multilayer Perceptron (MLP) with one or more fully connected layers, producing a probability distribution over the different output classes.



### 7.2.2 Computer vision-enriched discrete choice models

Throughout this paper, we consider the following choice situation. A decision-maker, $n$, faces a multi-attribute choice task with a set of $J$ mutually exclusive alternatives. Each alternative, $i$, is described by $M$ numeric attributes $X_i = \{x_{i1}, x_{i2}, ..., x_{iM}\}$, such as e.g. travel cost and travel time and by a (color) image $I_i$ with a resolution of $H \times W \times C$. The image captures attributes of the alternative, such as shape, form, or quality.

We assume decision-makers make decisions based on Random Utility Maximizing (RUM) principles [241], see Equation 7.1, where $U_{in}$ denotes the total indirect utility experienced by decision-maker $n$ considering alternative $i$, $V_{in}$ is the utility experienced by decision-maker $n$ derived from attributes observable by the analyst. And, to account for the fact that the analyst does not observe everything that matters to the decision-maker's utility, an additive error term $\varepsilon_{in}$ is added to each alternative [276].

$$U_{in} = V_{in} + \varepsilon_{in} \tag{7.1}$$

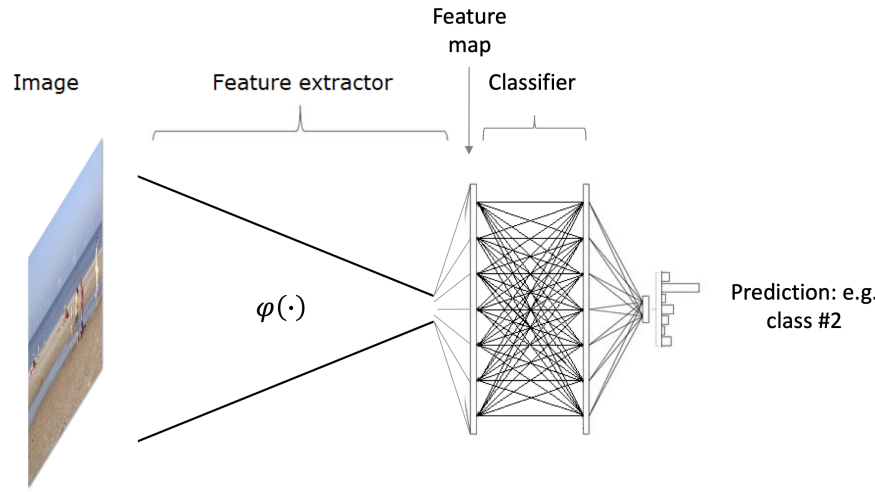



**Figure 7.1** Feature extraction and classification

Furthermore, we assume decision-makers experience utility from both the numeric attributes $X_i$ and the attributes encoded in the image $I_i$, see Equation 7.2, where $v$ is a preference function which maps the numeric attributes and the attributes encoded in the image of an alternative onto utility.



$$U_{in}(X_{in}, I_{in}) = v(X_{in}, I_{in}) + \varepsilon_{in} \tag{7.2}$$

In addition, we make three more assumptions to develop the CV-DCM:

1. We assume that the utility derived from the numeric attributes and the attributes encoded in the image are separable and additive in utility space, see Equation 7.3, where function $f$ maps the (observed) numeric attributes onto utility and function $g$ maps the attributes encoded in the image onto utility. Note that images typically encode multiple attributes. Therefore, the encoded attributes can be regarded as a composite good.

$$U_{in}(X_{in}, I_{in}) = f(X_{in}) + g(I_{in}) + \varepsilon_{in} \tag{7.3}$$

2. We assume that utility is linear and additive with numeric attributes as well as with the attributes encoded in the images, as captured in the feature maps. Thus, $f$ and $g$ are standard linear-additive utility functions. As discussed in section 2.1, feature maps are more compact representations of images. Accordingly, we let $Z_i = \{z_{i1}, z_{i2}, ..., z_{iK}\}$ denote the feature map of image $I_i$, and $\varphi(w) : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^K$ be a function that maps image $I_i$ onto feature

map $Z_i$. Hence, $\varphi$ is the transformation produced by the feature extractor of a CV model, and $w$ are its associated weights (i.e., the trainable parameters), which extracts the attributes encoded in the images. Both the numeric attributes $X_i$ and feature map $Z_i$ enter the utility function in a linear-additive fashion, as shown in Equation 7.4. In Equation 7.4, $\beta_m$ denotes the marginal utility associated with attribute $m$; $x_{imn}$ denotes the attribute level of numeric attribute $m$ of alternative $i$, as faced by decision-maker $n$; and $\beta_k$ denotes the weight associated with the $k$-th element of feature map $Z_{in}$.

$$U_{in} = \overbrace{\underbrace{\sum_m \beta_m x_{imn}}_{\substack{\text{Systematic utility}\\ \text{derived from}\\ \text{numeric attributes}}} + \underbrace{\sum_k \beta_k z_{ikn}}_{\substack{\text{Systematic utility}\\ \text{derived from attributes}\\ \text{encoded in the image}}}}^{v_{in}} + \varepsilon_{in}, \qquad \text{where } Z_{in} = \varphi(l_{in} \mid w) \tag{7.4}$$

The reason that we let feature maps, as opposed to individual pixel values, enter the (indirect) utility function is that letting pixels enter the utility function is at odds with the notion that utility is derived from consuming a certain bundle of goods and services. After all, pixels are not consumed; rather, utility is derived consuming a good, which is conceptualized in terms of their constituent attributes [268]. The feature map comprises the consumable attributes, encoded by the images' pixels. Thus, in the CV-DCM, $\varphi(w)$ produces a feature map containing the street-level attributes encoded in the image that are relevant to explain the choice behavior (and in such a way that they map linearly onto utility). However, it should be noted that its elements do not come with any a priori behavioral or semantic interpretation. The semantic meaning of the element may be extracted through post-hoc eXplanaible AI (XAI) analyses.

3. In line with common practice in choice modeling, we assume $\varepsilon_{in}$ is independent and identically Extreme Value Type I distributed with a variance of $\pi^2/6$, resulting in the well-known and convenient closed-form logit formula for the choice probabilities ($P_{in}$), given in Equation 7.5, where $C_n$ denotes the set of alternatives presented to decision maker $n$. Note that this assumption would, from a machine learning perspective, be equivalent to saying that the output layer is a Softmax function.

$$P_{in} = \frac{e^{v_{in}}}{\sum_{j \in C_n} e^{v_{jn}}} \tag{7.5}$$

Figure 7.2 depicts a graphical representation of the model structure of the proposed CV-DCM. It shows that the network's upper and lower parts are identical. This highlights an essential aspect of the CV-DCM's architecture: its consistency with RUM. It is consistent with RUM because it satisfies two conditions: regularity and transitivity (see

Hess et al. [256] for a rich discussion on RUM consistency and RUM consistency tests). This is evident from Equation 7.4, which shows that (1) the utility of one alternative does not depend on the attributes of another alternative, and (2) the utility function preserves ordinality. As a result, we can conceive the values at nodes in the last layer as utilities. However, even though we can interpret the last layer as utility, we cannot interpret $\beta_k$ in the same way we can with $\beta_m$. $\beta_k$ can be conceived as a marginal utility–after all, it reflects the change in utility by a unit change in the attribute level. But, because the meaning and units of the elements on the feature map, $Z_i$, are unclear, they do not carry a behavioral meaning. Furthermore, although it is technically possible (though challenging) to compute standard errors associated with $\beta_k$, this is not immediately a meaningful thing to do. After all, without interpretation of elements of the feature map, we do not have a hypothesis we wish to accept or reject. Having said that, in the situation in which meanings are attributed to certain elements of the feature map – e.g. through the use of XAI – computation of the standard errors could become meaningful.

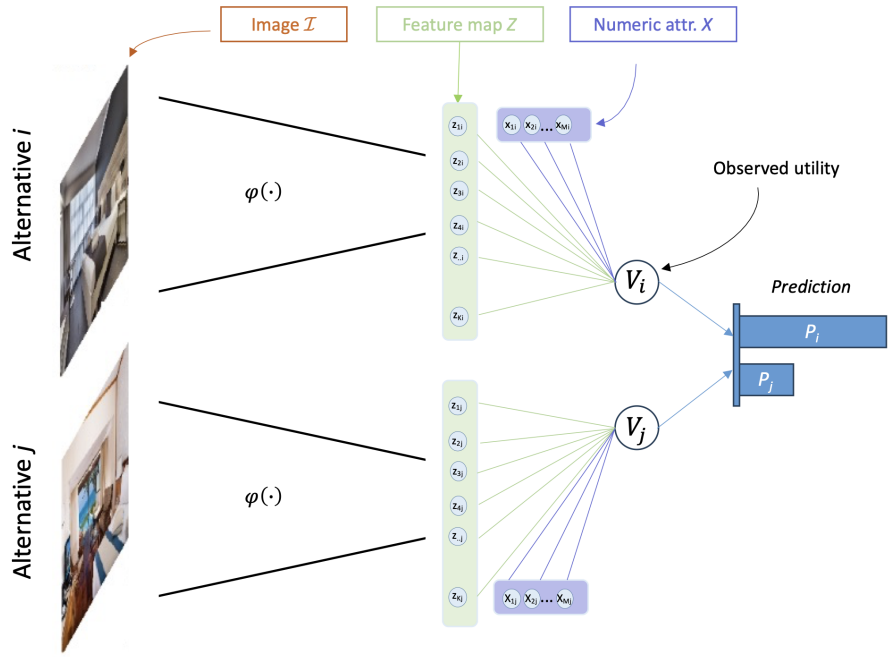



**Figure 7.2** Model structure of CV-DCM

### 7.2.3 Feature extractor and training

In this study, we use the feature extractor of the DeiT base model [277]. DeiT models are data-efficient vision transformer-based models that produce competitive capabilities on benchmark data sets, such as ImageNet [278], at a lower computational cost and data

requirements than many of its competitors [277]. The DeiT base model comprises a relatively modest 86 million weights and produces feature maps containing $K = 1,000$ elements. Furthermore, we use transfer learning to train our CV-DCM [279] to lower the computational time and amount of training data. The idea of transfer learning is to use a pre-trained network as the starting point for developing another network for a closely related task. In other words, rather than retraining the whole model from scratch, we start the training from an already good starting point when we train the CV-DCM. Our pre-trained DeiT base model is trained on ImageNet [280], a widely used benchmark image data set containing 1.2 million training images with 1,000 object classes.

## 7.3 Data collection and training

We demonstrate the proposed CV-DCM by applying it to data obtained through a stated choice experiment involving residential location choices. The residential location choice makes a suitable case study because both numeric attributes and street-level conditions, which we visualize using images, can be expected to be important to residential location choice behavior [281]. Street-level conditions can be thought of as a composite good that encompasses various elements, such as cleanliness, greenery, infrastructure quality, and overall aesthetic appeal. Moreover, images of the sort that we need for conducting a residential location choice experiment, namely street-level images, are widely available from map services such as Google, Apple, and Baidu and have been used in numerous scientific inquiries, including research on safety perceptions and people's density in urban places [22, 282–284]. Having access to a sufficiently large and diverse set of images is crucial for effectively training the feature extractor of the CV-DCM. While the exact number of images required is unknown before training, more images (and choice observations) generally lead to better training. In addition to their availability, street-level images have been shown to be a reliable representation of street-level conditions, as demonstrated by [285].



### 7.3.1 Stated choice experiment

In the Stated Choice (SC) experiment, we asked respondents to imagine they were required to move to a different neighborhood. They were presented with two alternatives for residential locations and asked to indicate which of the two they would choose. Figure 7.3 shows a screenshot of a choice task from the experiment. Prior to starting the choice experiment, respondents were provided with the following information:

1. Your new house is identical to your current house in terms of, e.g. size, type, built-year, furniture, maintenance, etc. Only your neighborhood changes.

2. Your monthly housing cost (including rent, mortgage, taxes, insurance, etc.) may go up or go down.

3. Your new neighborhood is relatively near your current neighborhood, but your commute time may still go up or down. The commute time is for your current mode of transport.

4. Your situation stays the same in all other aspects, e.g. in terms of distances to amenities, schools, the general practitioner, etc.

5. The images shown in the choice tasks depict the window view at ground level on the street side.

The alternatives comprise two salient numeric attributes: monthly housing costs ($hhc$) and commute travel time ($tti$). We choose these two attributes for three reasons. Firstly, they are known to be important to the residential location choice [286]. Secondly, they apply generically to almost everyone's residential location choice. Thirdly, they may help to interpret our empirical results. The combination of cost and time attributes allows us to compute the Value-of-Travel-Time (VTT), a metric that is widely studied in transport [287] and thus can be used for model validation. Finally, we did not include more attributes to the design because the paper's objective is to demonstrate the effectiveness of the proposed CV-DCMs to capture visual preferences instead of, e.g. developing a comprehensive model to predict residential location choices.

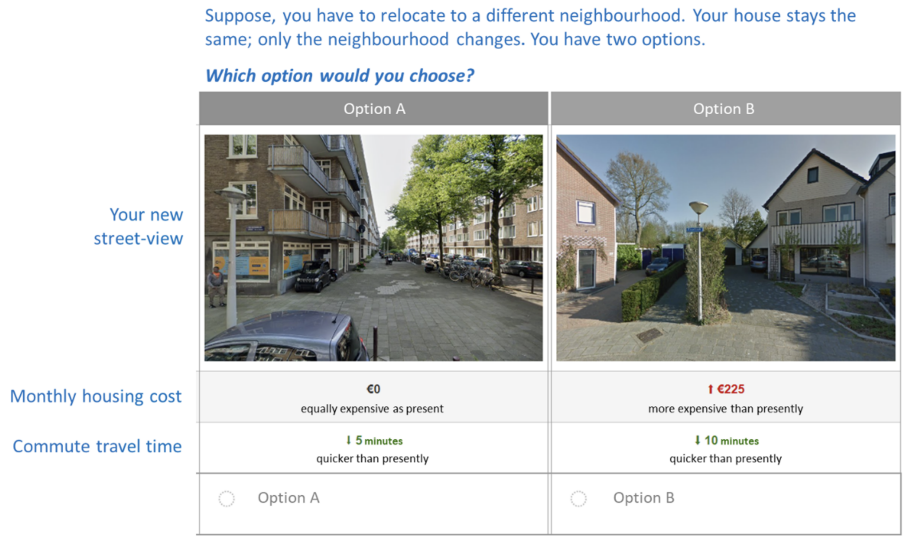



**Figure 7.3** Screenshot of the pivoted stated choice experiment (Image source: Google [231]) (translated to English; original in Dutch)

As can be seen in Figure 7.3, we have opted for a pivoted experimental design. We use a pivoted design to present respondents with as realistic choice situations as possible. Using absolute levels instead of pivoted levels would presumably render many

choice tasks unrealistic because of the considerable variation across respondents' current situations, especially regarding housing costs. For the attribute housing cost, we have used seven pivoted levels. For the attribute travel time, the number of levels and ranges we presented to the respondent depended on the respondent's current travel time, see Table 7.1. The ranges of both attributes were determined through a small pilot conducted before the actual survey.

**Table 7.1** Attribute levels Stated Choice experiment

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

### Street-level images

Besides monthly housing costs and commute travel time, each alternative comes with an image showing the street-level conditions. This image is randomly sampled from a database of street-level images we created before conducting the stated choice experiment. A major effort went into the construction of this database with street-level images. Specifically, we took the following steps to build the database. First, we randomly selected 50 municipalities (of about 350) in the Netherlands. We capped the number of municipalities to 50 because using more would lead to collecting many more images than we would need for our SC experiment. Second, we created a grid of points with 150-meter spacing within areas designated as residential areas (within the selected municipalities). Third, we retrieved the nearest street-view image id for each point on the grid using Google's API. We collected ids for all available images taken in 2020, or later. Each image id corresponds to a 360-degree panorama photo. Fourth, from each panorama, we generated two image urls with 90-degree angles to the direction of the street (to both directions). This latter ensures the images are 'window views' (e.g. as opposed to views parallel to the driving direction of the car taking the images). Finally, urls of images of poor quality were algorithmically removed. More specifically, urls to black images, blurred images and images with tilted horizons were removed. The final database contains the urls a little over 60k street-view images of residential streets from 50 municipalities in the Netherlands.

Importantly, for each image in our database, we also stored the month of the year in which the image was taken. The Netherlands lies in temperate zones, having four distinct seasons. Even though street-view images are usually collected on dry days, due

to the seasonality, street-view images taken in the winter may look different from those taken in summer. These differences might, in turn, impact the utility experienced by the respondent from the depicted local environment (and thus must be accounted for in our models).

**Experimental design**

We have used a random experimental design. Because the images do not possess ordinal or categorical levels, adopting an orthogonal or efficient experimental design strategy was not feasible, at least not considering the images. Therefore, we took a two-step approach to construct the choice tasks. First, we randomly pulled a pair of images from our image database. The only requirement imposed on the drawing was that the drawn images were not from the municipality where the respondent lives. We determined each respondent's municipality (and province) based on the postcode we elicited at the start of the survey. We excluded images from the respondents' municipalities to avoid unobserved heterogeneity entering our experiment, which may be derived from respondents' knowledge of places where the images were taken. Unobserved utilities flowing into stated choice experiments could lead to biased modeling outcomes if not econometrically accounted for (see, e.g., Train and Wilson [288]; Van Cranenburgh et al. [289]; Guevara and Hess [290]). While excluding images from respondents' own municipalities does not guarantee that respondents do not recognize the places the street-view images were taken, it lowers the probability.

Second, we added the housing cost ($hhc$) and travel time ($tti$) levels. To do so, we randomly pulled a choice task from one of three tables with choice tasks we generated before conducting the SC experiment. Each table was created by taking the following steps. First, a full-factorial design was created based on the attribute levels shown in Table 1. Second, we excluded choice tasks that did not involve a trade-off between housing costs and travel time. Removing such (partially) dominating choice tasks is possible because we have strong prior beliefs for the expected sign of the preference parameters for housing cost and travel time. Third, we excluded all choice tasks where one or more attribute levels were equal. As a result of this choice task construction approach, each choice task necessarily consists of a trade-off between housing cost and travel time.



## 7.3.2 Data collection and sample description

The survey was implemented in SurveyEngine software and conducted in September 2022. The survey started with a few questions to determine respondents' eligibility for the survey. In particular, we elicited respondents' age, gender, postcode, and current commute travel time. Then came the SC experiment, in which each respondent was presented with 15 choice tasks. The images used in the choice tasks were directly retrieved from Google servers based on the urls from our image database. The survey ended with a series of questions regarding the respondents' current housing situation (e.g. housing costs, rating of the current visual street-level conditions) and commute

situation (e.g. mode of transport, number of commute days). Noteworthy, we also asked respondents how important the three attributes (housing cost, travel time and street-level conditions) were for their decisions on a scale from 1 to 10. Although it is well-known that direct elicitation of preferences is treacherous [291], it still can provide first (albeit inconclusive) evidence of the importance of the street-level conditions, presented using the images, relative to the numeric attributes for the residential location choices.

The target population for the survey was the Dutch population of 18 years and older, with ten or more minutes of commute travel time. The latter requirement was necessary because we used a pivoted experimental design. Because of this latter condition, no official population statistics exist to compare our sample against, but we do not expect this condition to affect the population statistics substantially. Therefore, care was taken in that the sample was, by and large, representative of the Dutch 18-year-old and older population in terms of gender, age, and spatial distribution across the Netherlands. Cint[1], a panel data provider, provided the panel of respondents. In total, 800 respondents completed our survey.

Table 7.2 shows the sample statistics. Overall, the sample is representative of the target population. Also, for the variables that are not explicitly considered during the data collection, such as the modal split and household composition, the statistics are close to the population data (c.f. Ton et al. [292]). Furthermore, looking at the reported monthly housing cost, we notice that the largest share of the respondents has a housing cost below 750. This seems reasonable since the average net housing cost of rental houses in the Netherlands is around 700 p/m; homeowners' average net housing cost is slightly above 900 p/m [293].

[1] see www.cint.com

**Table 7.2** Sample statistics

| Socio-demographic variable | Category | Distribution |
|---|---|---|
| **Age** | 18–29 year ("young") | 21% |
| | 30–39 year ("young") | 19% |
| | 40–49 year ("middle") | 20% |
| | 50–59 year ("middle") | 22% |
| | 60–69 year ("old") | 17% |
| | 70+ year ("old") | 1% |
| **Gender** | Male | 50% |
| | Female | 50% |
| **Province** | North (Groningen, Friesland, Drenthe) | 12% |
| | East (Gelderland, Overijssel) | 23% |
| | South (Limburg, Noord-Brabant, Zeeland) | 24% |
| | West (N-Holland, Z-Holland, Utrecht, Flevoland) | 41% |
| **Current commute travel time (TT)** | 10 min $< TT <$ 20 min | 35% |
| | 20 min $< TT <$ 30 min | 31% |
| | 30 min $< TT <$ 45 min | 20% |
| | 45 min $< TT$ | 14% |
| **Primary mode for commute** | Bike, E-bike, Scooter, Moped | 30% |
| | Bus, Metro, Tram | 8% |
| | Train | 10% |
| | Car, Motor bike | 52% |
| **Commuting days per week** | 1 day per week | 8% |
| | 2 days per week | 15% |
| | 3 days per week | 20% |
| | 4 days per week | 22% |
| | 5 or more days per week | 35% |
| **Household composition** | One-person household | 26% |
| | Multiple-person household without children | 40% |
| | Multiple-person household with children | 34% |
| **House type** | Flat, gallery, porch, apartment | 23% |
| | Terraced house | 31% |
| | Corner house | 16% |
| | Semidetached house | 14% |
| | Detached house | 15% |
| **Current monthly housing cost (HC)** | $HC <$ 750 p/m | 36% |
| | 750 $< HC <$ 1,250 p/m | 33% |
| | 1,250 $< HC <$ 1,750 p/m | 16% |
| | 1,750 $< HC$ p/m | 6% |
| | I do not want to report | 9% |
| **Rating of own visual street-level conditions** | 1 (worst) | 1% |
| | 2 | 6% |
| | 3 | 21% |
| | 4 | 46% |
| | 5 (best) | 26% |

### 7.3.3 Descriptive analysis

Figure 7.4 shows histograms of the self-reported importance levels of the street-level conditions (left), monthly housing costs (middle) and commute travel times (right). Figure 4 shows that the street-level conditions and monthly housing costs are, on average, considered equally important to the residential location choice and more important than commute travel times. The variance in the ratings across respondents is higher for the street-level conditions than for the monthly housing cost—suggesting a considerable amount of preference heterogeneity is present in the importance of street-level conditions. However, we observe the highest variance for the commute travel time. Noteworthily, the importance rating for the street-level conditions is weakly negatively correlated with the importance ratings for monthly housing costs ($\rho = -0.10$) and uncorrelated with the ratings for commute travel time ($\rho = 0.02$). In contrast, the importance ratings for monthly housing costs and commute travel times are strongly positively correlated ($\rho = 0.36$). This strong positive correlation reveals that people who find housing costs important usually also find commute travel time important, and vice versa.

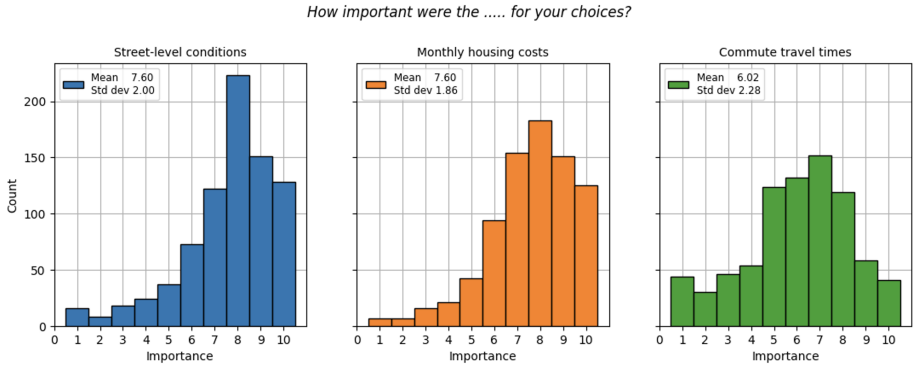



**Figure 7.4** Self-reported importance levels of attributes in the SC experiment

Figure 5 shows the Pearson correlation coefficients between importance ratings and a selection of respondent characteristics. Interestingly, the top row shows that the importance of the street-level conditions correlates strongest with the self-reported rating of respondents' current visual street-level conditions. This strong positive correlation suggests that people living in visually attractive neighborhoods consider their visual street-level conditions relatively more important than people living in visually less attractive places. This observation aligns with [294], who also find that the current situation affects residential location choice behavior. Moreover, we see that the importance of the street-level conditions positively correlates with living in a detached or semi-detached house. A self-selection mechanism could explain this effect: people caring about their visual street-level conditions are more likely to choose an attractive residential location (see e.g., Van Wee [295] for discussions on self-selection effects in

residential location choices; [296]). Finally, perhaps somewhat counter to expectations, we see that variables such as gender and monthly housing costs do not strongly correlate with the importance given to street-level conditions.

Furthermore, Figure 7.5 reveals that the importance of the monthly housing cost (middle row) correlates strongest with living in house type 'Flat, gallery, porch, or apartment'. This correlation seems in line with intuition, given that low-income people are more likely to live in this type of housing. Finally, we see that the importance of the commute travel time (bottom row) positively correlates with age class 18-39 years. Since this age class sits in the center of the working-age population, it makes sense that commute travel time is essential to this group. Altogether, the correlations reported in Figure 7.5 seem plausible.

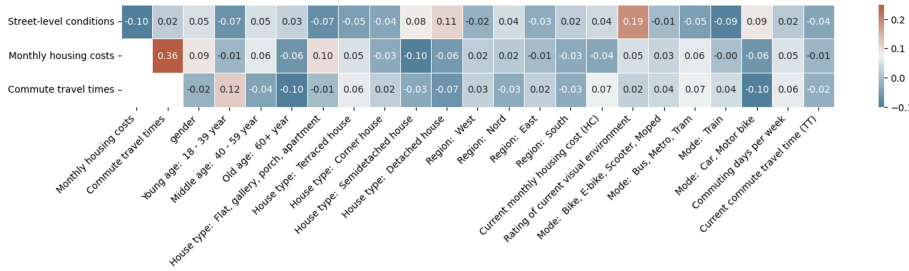



**Figure 7.5** Pearson correlation coefficients between importance ratings and respondent characteristics

7

Next, we analyze the images used in the stated choice experiment. Although our street-view image database comprises urls to over 60k images, only slightly over 7.5k unique images are used in the stated choice experiment. Because images are drawn randomly from our image database with replacement, we expect that some images will be sampled more than once. Indeed, most images are used once. However, contrary to our design intentions, some images are used 20 times or more. A possible underlying cause could be the seed numbers used by the survey platform's software. Nevertheless, regardless of this issue's origin, when we deal with the issue carefully during the training of our models (see below Train-test split), it does not need to have an impact on our (substantive) findings.

Finally, Figure 7.6 shows the distribution of the month of the year of the images used in the survey. In line with expectations, the images are not evenly distributed over the year. We see that most images are taken in spring and summer (March to September). Furthermore, we notice that images have been sampled for all 12 months. This implies we can account for the impact of the seasons on the utility derived from the street-view images by estimating constants for all months (except one, which we need to fix to zero for normalization).

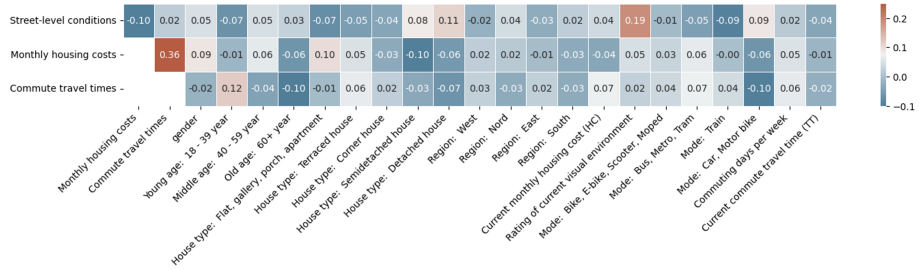



**Figure 7.6** Distribution of images used in the stated choice experiment over the months of the year.

## 7.3.4 Training

### Loss function and implementation

Training a CV-DCM involves finding the weights of the model ($\beta$, $w$) that minimize the loss function. In other words, the weights of the feature extractor and preference parameters of the utility function are jointly optimized. For this study, we use a cross-entropy loss function with an L2 regularization term, see Equation 7.6. Minimizing the cross-entropy loss is equivalent to maximizing the Log-Likelihood of the data given the model—which is common practice in the choice modeling literature. The L2 regularization aims to reduce the chance of model overfitting by penalizing the magnitude of the weights in the model. $\gamma$ governs the strength of the regularization. Note that we apply regularization only to $w$ and not to preference parameters $\beta_m$ and $\beta_k$. Regularizing preference parameters could lead to undesirable biases.

$$\hat{\mathbf{w}},\, \hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{w},\,\boldsymbol{\beta}} \left[ \overbrace{-\frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{J} y_{nj}\, \log\big(P_{nj} \mid X_{nj},\, l_{nj},\, \boldsymbol{\beta},\, \mathbf{w}\big)}^{\text{cross–entropy loss}} + \overbrace{\gamma \sum_{r} w_r^2}^{L_2 \text{ regularisation}} \right]. \tag{7.6}$$

We have made the data openly available[2]. By doing so, we aim to support model-building and validation practices. We hope our data can become a benchmark data set for studying choice behavior in the presence of visual stimuli.

### Implementation and hyperparameter tuning

Our CV-DCM is implemented and trained in PyTorch [297]. PyTorch is a Python-based machine learning package commonly used for deep learning computer vision research because it supports GPU computing. We conducted hyperparameter tuning, in which we performed a grid-search over the most important hyperparameter: the optimization

2

algorithm, learning rate, batch size, and L2 regularization (see Table 7.3). All other (hyper)parameters (such as dropout rates, layers and activation functions) were kept at their default values.

**Table 7.3** Hyperparameter tuning CV-DCM

| Hyperparameter | Hyperparameter space |
|---|---|
| Optimisation algorithm | {Adam, SGD} |
| Batch size | {12, 16, 20, 24} |
| L2 weight decay ($\gamma$) | {0, 0.1, 0.2, 0.3} |
| Learning rate | $\{1e^{-5}, 1e^{-6}\}$ |

**Image transformation and feature scaling**

In line with common practice in computer vision, we transform and augment images while training the CV-DCM. Specifically, we conduct the two operations. First, we downsampled the images to $224x224$ pixels. This downsizing operation ensures that images have the input dimensions expected by the CV model (i.e. DeiT base model). Second, we randomly flip images horizontally. This data augmentation operation reduces the model's ability to remember images, thus lowering the chance of overfitting the training data. Furthermore, we scale the numeric features. Scaling the features helps the optimizer to avoid getting stuck in local minima [298]. The most common type of scaling in machine learning involves shifting and scaling the features to a zero mean and a unit variance. We use another commonly used scaling technique to scale the housing cost and travel time features, called min-max scaling. This scaling entails scaling the features to a range of $[-1, 1]$. The advantage of this scaling technique is that it is straightforward and facilitates easy interpretation of the model's parameters. To facilitate interpretation, we have used the same scaling for all data (thus ignoring that the minimum travel time level varied across respondents, see Table 7.1). Specifically, all housing costs are divided by 225 and travel times are divided by 15.

**Train-test split**

Splitting the data into a train set and a test set is essential for training virtually all machine learning models because their high capacity makes them prone to overfitting [298]. As the name suggests, the train set is used for training the model; the test is unseen by the model during training and used to evaluate (test) the model's generalization performance after training. If a trained model overfits the data, a gap in the performance between the train and test set will tell.

The most common way to create the train-test split is by randomly allocating observations to the two sets. When splitting data, it is important to avoid "data leakage". Data leakage happens when the model has access to information during training that it does not have when deployed after training (see e.g. Hillel [299] for its impact on choice model outcomes). For this study, we split our data across images. Thereby, we aim to

avoid potential data leakage from learning the utility levels of specific images rather than generalizable high-level utility-generating features embedded in the images. Making such a split is, however, a nontrivial network problem. Every image is connected at least to one other image (the other street-view image presented in the choice task). However, some images are connected to dozens of other images because they are used more than once (see section 7.3.3). Hence, when we assign one image to the train set, we must also place all directly and indirectly connected images in the train data set.

Given the above 'network' problem, we followed the following procedure to create the train and test sets. We randomly picked one choice task, comprising two images, and put this choice task and all choice tasks connected to this one in the train set. We repeated the random picking of choice tasks until 80% of the data were used. The remaining data (20%) make the test data set. The train and test data sets comprise, respectively, $N = 9,784$ and $N = 1,948$ choice observations. Due to our splitting strategy, observations of the same individual may be present in both the train and test data sets. However, it is unlikely to cause serious data leakage because no socio-demographic variables (that would be needed to identify observations of the same respondent) are used in the training of the CV-DCM.

## 7.4 Results

We estimate/train four models on the residential location choice data whose utility functions are given in Equation 7.7 to Equation 7.10. Models 1 and 2 are standard linear-additive RUM-MNL models used as benchmark models to compare the proposed CV-DCM (Model 3). Model 1 ignores the images completely, while Model 2 takes into account the month in which the image is taken by estimating constants, denoted $\beta_{mo}$, for each month. If where and when images are collected are uncorrelated, we expect that images taken in spring and summer, on average, attain a higher utility than images taken in autumn or winter. Model 3 is the proposed CV-DCM and takes the monthly housing cost ($hhc$), commute travel time ($tti$), and the month of the year as numeric input attributes in the same way as Model 2 does, but also takes the feature maps of the images as inputs. Finally, Model 4 is similar to Model 3 but interacts the feature map with age group. Thereby, this model is able to capture systematic taste heterogeneity across age groups over the images.



$$\text{Model 1} \quad U_{in} = \beta_{hhc}\, hhc_{in} + \beta_{tti}\, tti_{in} + \varepsilon_{in} \tag{7.7}$$

$$\text{Model 2} \quad U_{in} = \beta_{hhc}\, hhc_{in} + \beta_{tti}\, tti_{in} + \sum_{mo} \beta_{mo}\, I_{in} + \varepsilon_{in} \tag{7.8}$$

$$\text{Model 3} \quad U_{in} = \beta_{hhc}\, hhc_{in} + \beta_{tti}\, tti_{in} + \sum_{mo} \beta_{mo}\, I_{in} + \sum_{k} \beta_{k}\, z_{ikn} + \varepsilon_{in} \tag{7.9}$$

$$\text{Model 4} \quad U_{in} = \beta_{hhc}\, hhc_{in} + \beta_{tti}\, tti_{in} + \sum_{mo} \beta_{mo}\, I_{in} + \sum_{age} \sum_{k} \beta_k^{age}\, age\, z_{ikn} + \varepsilon_{in}$$

$$\text{where} \quad I_{in} := \begin{cases} 1 & \text{if } mo = l_{in}^{mo}, \\ 0 & \text{else} \end{cases}$$

$$Z_{in} = \varphi(l_{in} \mid w),$$
$$\varepsilon_{in} \sim i.i.d. \text{ Extreme Value Type I.} \tag{7.10}$$

### 7.4.1 Estimation results

Table 7.4 shows the estimates for the behavioral interpretable parameters as well as the model performance on the train and test sets, using three (related) metrics: the Log-Likelihood, rho-square and cross-entropy. A good model performance on the test set implies the model generalities well to new/unseen data. It is important to note that we compare the model performance of the CV-DCMs with Models 1 and 2 to get a feeling of how much of the unexplained variance is captured by adding the computer vision model. Models 1 and 2 are not meant as a yardstick to show that the CV-DCMs outperform them. Because models 1 and 2 do not account for the images, they are unlikely to be used in practice on these data. In this regard, a more even-handed comparison would be models that first encode information in the images in tabular form, such as done by Ramirez et al. [51]. However, such a comparison would involve a study in itself and thus go beyond the scope of this paper.



We can draw three conclusions based on the performance metrics in Table 7.4. The first and most important conclusion is that the CV-DCM can extract relevant information from the street-level images to predict the choice behavior. Looking at the generalization performance, we see that the plain vanilla CV-DCM (Model 3) outperforms the two benchmark models (Model 1 and 2) by a fair margin. Specifically, the CV-DCM improves the Log-Likelihood on the test set by 57 Log-Likelihood points, and the rho-square jumps from $0.116$ to $0.158$[3]. Since Model 3 collapses into Model 2 when setting $\beta_k = 0 \forall k$ (see Equations 7.8 and 7.9), we can statistically compare their model fit using a Likelihood Ratio Statistic (LRS). The LRS exceeds far the critical level of significance (set at $\alpha = 0.05$), with $K = 1,000$ degrees of freedom, supporting the notion that the CV-DCM's capability to handle images leads to a statistically significant improvement in model fit. Second, the month of the year carries limited information regarding the utility generated by the images, at least when used in isolation from other information from the images, as in Model 2. Comparing Models 1 and 2, we observe that Model 2

[3]Note that the rho-square on the test set is slightly higher than the rho-square on the training set. This is presumably caused by small differences between the training and test sets. For instance, some observations that are relatively poorly explained by Model 3 may have ended up in the training set by coincidence. This, in turn, causes the rho-square of the training set to be relatively worse than the rho-square on the test set.

outperforms Model 1 by 23 Log-Likelihood points on the train set but performs on par on the test set. Hence, the incorporation of the month of the year in the utility function does not improve the generalizability of the conventional RUM-MNL models. Third, comparing Models 3 and 4, we see that allowing for an interaction between the feature map and age category (young, middle, old) further improves the model's generalization performance. This reveals the presence of systematic taste heterogeneity concerning street view conditions across age groups.

Despite having to train 86 million weights, the extra computational time does not render the CV-DCM impractical; 1.5 hours of training time is in the same order of magnitude as the estimation time of advanced mixed logit models. Having said that, handling large numbers of images and working with GPUs is technically considerably more challenging than estimating a conventional discrete choice model using an off-the-shelf estimation package. Moreover, deriving the standard errors for the CV-DCM can be more demanding. To obtain the standard errors for $\beta_m$ reported in Table 7.4, we re-estimated Model 3 and Model 4 while fixing the utilities derived from the attributes encoded in the images. This approach is straightforward but not helpful when a researcher wants to derive the standard errors associated with $\beta_k$ (and suboptimal when attributes encoded in the images are correlated with numeric attributes). Computing the standard errors associated with $\beta_k$ for Model 3 turns out to be computationally demanding and technically challenging (because of the large number of estimates and collinearity).

Next, we look at the estimated taste parameters. We see that housing cost and commute travel time are highly relevant attributes to the residential location choice. In line with expectations, $\beta_{tti}$ and $\beta_{hhc}$ are highly significant, and their minus signs align with behavioral intuition. Based on $\beta_{tti}$ and $\beta_{hhc}$, we also compute the VTT. In the context of our SC experiment, the VTT gives the (mean) willingness to pay per month for a one-hour travel time reduction per commute trip. A VTT between €217 and 228 per hour per month seems reasonable, considering that most respondents in our sample commute five days per week, and thus about 20 days per month. Furthermore, in line with expectations, we observe that the VTT is stable across all models. We expect stable $\beta_{tti}/\beta_{hhc}$ ratios because our experimental design is constructed in such a way that images and numeric attribute levels within choice tasks are entirely uncorrelated. Cramer [300] shows that ratios of logit model estimates are unaffected by omitted variables if the omitted variables are uncorrelated with other explanatory variables.

The signs of the estimates associated with the months of the year are mostly intuitive. These estimates reflect the average utility difference between an image taken in that month and images taken in December (which we fixed to zero). In Model 2, we see that the estimates associated with the months of the year are mostly positive and significant for the spring and summer months. This can be explained by the notion that images taken in these months are more likely to look more attractive to live than images taken in winter, for instance, because the weather is better. However, the positive and significant estimate for January counters this line of argumentation and is hard to explain. In Models 3 and 4, the estimates associated with the months of the year do not carry the same interpretation as under Model 2. The utility derived from an image in Models 3 and 4 is the sum of

the utility from the image's feature map and the estimate associated with the month of the year of the image. As a result, we cannot see the estimates associated with these two utility sources in isolation. One noteworthy observation concerning the estimates related to the months of the year in Models 3 and 4 is that fewer estimates are significant than in Model 2. This observation aligns with statistical expectations. Because feature maps already contain information about the weather conditions in the month of the year, explicitly adding the month to the model provides comparatively less information to explain the choice behavior. For example, an image in which trees that have shed their leaves reveals the image is probably taken in winter. See Sifringer and Alahi [301] for recent work on handling data congruency.

**Table 7.4** Estimation results

| | | **Model 1** | | | **Model 2** | | | **Model 3**[IV] | | | **Model 4**[V] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model type | | lin-add RUM-MNL | | | lin-add RUM-MNL | | | CV-DCM | | | CV-DCM with interaction | | |
| Number of parameters | | 2 | | | 13 | | | 86m | | | 86m | | |
| Estimation time | | <1 sec[I] | | | <1 sec[I] | | | 1.5 hr.[II] | | | 1.5 hr.[II] | | |
| Train set N = 9,784 | Log-Likelihood | -5,954 | | | -5,931 | | | -5,724 | | | -5,304 | | |
| | $\rho^2$ | 0.120 | | | 0.130 | | | 0.156 | | | 0.218 | | |
| | Cross-entropy | 0.609 | | | 0.606 | | | 0.585 | | | 0.542 | | |
| | Hit-rate (accuracy) | 0.695 | | | 0.697 | | | 0.716 | | | 0.748 | | |
| Test set N = 1,948 | Log-Likelihood | -1,194 | | | -1,194 | | | -1,137 | | | -1,119 | | |
| | $\rho^2$ | 0.116 | | | 0.116 | | | 0.158 | | | 0.171 | | |
| | Cross-entropy | 0.613 | | | 0.613 | | | 0.585 | | | 0.574 | | |
| | Hit-rate (accuracy) | 0.690 | | | 0.687 | | | 0.697 | | | 0.710 | | |
| | | *est* | *s.e.* | *p-val* | *est* | *s.e.* | *p-val* | *est* | *s.e.*[III] | *p-val*[III] | *est* | *s.e.*[III] | *p-val*[III] |
| $\beta_{hhc}$ | | -0.86 | 0.025 | 0.00 | -0.87 | 0.024 | 0.00 | -0.96 | 0.025 | 0.00 | -0.93 | 0.025 | 0.00 |
| $\beta_{tti}$ | | -0.21 | 0.023 | 0.00 | -0.21 | 0.025 | 0.00 | -0.24 | 0.026 | 0.00 | -0.23 | 0.026 | 0.00 |
| $\beta_{jan}$ | | | | | 0.46 | 0.129 | 0.00 | 0.25 | 0.136 | 0.07 | -0.02 | 0.137 | 0.86 |
| $\beta_{feb}$ | | | | | 0.02 | 0.228 | 0.91 | -0.40 | 0.240 | 0.10 | 0.02 | 0.242 | 0.92 |
| $\beta_{mar}$ | | | | | 0.10 | 0.080 | 0.23 | 0.05 | 0.084 | 0.58 | -0.04 | 0.084 | 0.63 |
| $\beta_{apr}$ | | | | | 0.25 | 0.080 | 0.00 | 0.36 | 0.084 | 0.00 | 0.04 | 0.085 | 0.66 |
| $\beta_{may}$ | | | | | 0.28 | 0.084 | 0.00 | 0.08 | 0.088 | 0.39 | 0.01 | 0.089 | 0.89 |
| $\beta_{jun}$ | | | | | 0.17 | 0.084 | 0.04 | -0.12 | 0.088 | 0.16 | 0.01 | 0.088 | 0.95 |
| $\beta_{jul}$ | | | | | 0.21 | 0.094 | 0.02 | 0.31 | 0.098 | 0.00 | -0.11 | 0.099 | 0.26 |
| $\beta_{aug}$ | | | | | 0.24 | 0.087 | 0.01 | 0.12 | 0.092 | 0.17 | -0.02 | 0.092 | 0.82 |
| $\beta_{sep}$ | | | | | 0.19 | 0.085 | 0.03 | 0.33 | 0.089 | 0.00 | -0.07 | 0.090 | 0.46 |
| $\beta_{oct}$ | | | | | 0.46 | 0.131 | 0.00 | 0.40 | 0.138 | 0.00 | -0.10 | 0.138 | 0.47 |
| $\beta_{nov}$ | | | | | -0.11 | 0.106 | 0.31 | -0.04 | 0.111 | 0.74 | -0.19 | 0.111 | 0.08 |
| $\beta_{dec}$ | | | | | 0.00 | --fixed | | 0.00 | --fixed | | 0.00 | --fixed | |
| Value-of-Travel-Time [€/hr month] | | 216.7 | 28.26 | 0.00 | 217.2 | 28.35 | 0.00 | 228.5 | 26.73 | 0.00 | 225.7 | 26.08 | 0.00 |

[I] Using 4 CPUs (Xeon @ 3.60 GHz)
[II] Using GPU (GeForce RTX 2080Ti)
[III] Obtained through computing the hessian while keeping the utility derived from the image fixed
[IV] Optimal hyperparameters: {optimiser: SGD, Batch size: 20, L2: 0.1, Learning rate: 1e-6}
[V] Optimal hyperparameters: {optimiser: SGD, Batch size: 24, L2: 0.1, Learning rate: 1e-5}

Lastly, we analyse the contributions to utility differences between the right and left-hand side alternatives derived from the images' feature maps. To do so, Figure 7.7 shows three kernel density plots for the plain vanilla CV-DCM (Model 3). The left-hand side plot shows the total utility difference as predicted by the trained CV-DCM; the middle plot shows the utility difference from the numeric attributes; and the right-hand side plot shows the utility difference from the attributes encoded in the images. We make several observations based on Figure 7.7. Firstly, looking at the range of x-axes of the middle and right-hand side plots, we see that the utility differences arising from the

street-level conditions and numeric attributes are similar. This tells us that the part-worth utilities derived from the numeric attributes (housing cost and travel time) are of the same magnitude as those derived from the street-level conditions embedded in the street-view images. This observation adds to the evidence that street-level conditions are important to residential location choice behavior and can effectively be modeled using images and CV-DCMs. Secondly, we notice that the distributions of utility differences are virtually equal for the test and train sets. This indicates that the CV-DCM does not overfit the training data, and the data are adequately split into train and test sets. Therefore, the CV-DCM must have learned to extract salient generalizable features from the images that generate utility. Thirdly, we see that the distribution of the utility differences stemming from the images is comparatively more bell-shaped than those of the numeric attributes. At first sight, this may seem odd, but it can be explained by how the choice tasks have been constructed. Recall that we removed choice tasks without trade-offs between the numeric attributes (see section 7.3.1 for more details). This removal leads to the bi-modal shape of the utility difference.

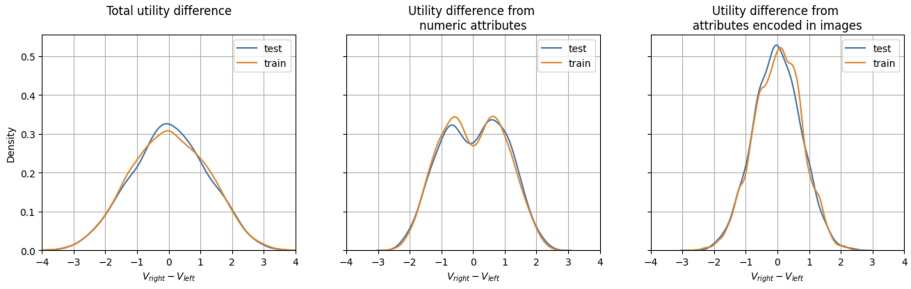



**Figure 7.7** Utility differences.



### 7.4.2 Face validity: what has the CV-DCM learned about street-level conditions?

The improvement in model performance by the CV-DCMs compared to the benchmark models supports the notion that the CV-DCMs can extract relevant information from images to predict choice behavior. But, $\beta_k$ and $w$ do not carry a behavioral meaning. Therefore, they do not provide directly interpretable insights about what the CV-DCMs have learned regarding what decision-makers find important for their residential location choices. To shed light on what the CV-DCM has learned about the decision-makers' preferences, we show two collages of images taken from the test set, to which the trained CV-DCM (Model 3) assigns the highest (Figure 7.8) and lowest (Figure 7.9) utility levels. Note that the utility level is stamped in the top left of each image. These utility levels are 'uncorrected' for the month of the year. Hence, the top left image yields a utility of $1.63$ if the image was taken in December, while it produces a utility of $1.63 - 0.12 = 1.51$ if it was taken in August (which it is).

What catches the eye in Figure 7.8 is that the images all look spacious, leafy and often water-abundant. We see many trees, grassland and detached houses. In the authors' view, these street-level conditions are indeed highly attractive. In sharp contrast, the images in Figure 7.9 look cramped, grayish, and urbanized and often have hallmarks of transportation, such as overhead wires, bus stops, parked bikes, and cars. In the authors' view, these street-level conditions are indeed highly unattractive. The mean difference in utility between the 20 best, shown in Figure 7.8, and the 20 worst street-level conditions, shown in Figure 7.9, is 2.7 utility points. The willingness to pay per month to move from the worst to the best street-level conditions can be computed by dividing the utility difference by $\beta_{hhc}$. The result yields a willingness to pay of 632 euros per month – which seems high but perhaps not implausible. Here, it should be noted that this estimate concerns the two most extreme street-level conditions.


**Figure 7.8** Images showing street-level conditions with the highest predicted utility levels (based on the plain vanilla CV-DCM)

## 7.4.3 Policy-relevant insights

After establishing the CV-DCM approach's face validity, we can use it to obtain policy-relevant insights. Given that the paper's main objective is methodological, we present only two brief examples.

### Effect of age on preferences over street-level conditions

A policy-relevant insight that can be gleaned from the CV-DCM is the effect of age on preferences over street-level conditions. It is well established that different age generations have different housing needs (e.g. Booi et al. [302]). As such, it seems plausible that they also have different preferences over street-level conditions. To develop

**Figure 7.9** Images showing street-level conditions with the lowest predicted utility levels (based on the plain vanilla CV-DCM)

new housing policies targeted at specific age generations, a thorough understanding of what street view conditions are considered attractive by which age generation is required. For this analysis, we use the trained CV-DCM with age interactions (Model 4). Using this model, we computed the utilities for each of the $\sim 7.6k$ images used in the SC experiment for young and old people. Then, we look at the extent to which the utilities correlate between young and old people and, more interestingly, where they deviate.

Figures 7.10 and 7.11 show the results. More specifically, Figure 7.10 shows sixteen street views that are comparatively more attractive to young people than older people; Figure 7.11 shows sixteen street views that are comparatively more attractive to older people than young people. At the top left in each figure, the utility levels predicted by the CV-DCM for young and old people are shown. Furthermore, a kernel density plot shows the part-worth utility distribution (top) on the right-hand side of each figure. The vertical lines in this plot indicate where the sixteen depicted images sit in the overall distribution for younger (blue) and older (orange) people. A scatter plot scatters the part-worth utilities to young (x-axis) and old (y-axis) people at the bottom of the right-hand side plot. The red dots in the scatter plots correspond to the depicted street views.

Based on Figures 7.10 and 7.11, a couple of policy observations can be made. Firstly, the scatter plots show a moderate correlation between the part-worth utilities from street-level conditions to younger and older people ($\rho = 0.76$). This means that, across the board, young and old people tend to agree on what attractive and unattractive street-level conditions are. But there are also street-level conditions where the utilities clearly diverge. In particular, Figure 7.10 shows that young people find suburban areas relatively more attractive than old people. Most images in Figure 7.11 show hallmarks of

suburbia, like terraced houses, parking facilities, gardens, (boutique) shops, and streets. Likewise, Figure 11 shows that older people find greener, leafier areas with fewer houses and cars comparatively more attractive than young people. These results are in line with general beliefs. From a policy perspective, they underpin the need to consider the preferences over street-level conditions of the target population when developing new housing projects.


**Figure 7.10** Images showing street-level conditions that are comparatively attractive to younger people as compared to older people

7

### Relationship between visual attractiveness and population density

Faced with scarcity of land and increasing population levels, various Western European governments have developed housing policies with the aim of creating compact, high-density cities (e.g., by building more high-rises). Previous research, however, suggests that low-density (rural) areas are considered to be more visually appealing and scenic [303] and that this heightened visual attractiveness is one of the main motivations for "counter-urbanism" [304], which is characterized by people moving away from urban areas and settling in rural or suburban areas. Counter-urbanizing could thus undermine policies designed to create compact cities and put more strain on already burdened transportation networks. Previous studies into counter-urbanism have mostly relied on proxies for visual attractiveness, such as shares of older housing, proximity to natural areas, and the number of nearby hotels.

The trained CV-DCMs allow a more direct examination of the relationship between population density and visual attractiveness of the street-level conditions. To do so, we merge the population density at the location where the image is taken onto our image data set containing $\sim 7.6k$ randomly sampled street view images from the Netherlands. Then, we use Model 3 to compute the utility level for each image. Finally, we group the images based on population density quantiles.


**Figure 7.11** Images showing street-level conditions that are comparatively attractive to older people as compared to younger people

Figure 7.12 presents the results of this analysis in a box plot. In line with the motivation for counter-urbanism, we find evidence that low-density (rural) areas have more attractive street-level conditions. In fact, using t-tests, we have established that the mean values of each population density quantile are significantly distinct from the means of all the other quantiles. This implies that there are notable differences in the utility of street-level conditions across levels of population density. From a policy perspective, these results suggest that policies aimed at creating compact cities should seriously take the attractiveness of street-level conditions into account. Although this study does not investigate counter-urbanism, failing to do so may undermine the effectiveness of such policies, as they may push people towards the suburbs and beyond.



## 7.5 Conclusion and discussion

This paper contributes to the recent methodological progress made in the fields of transportation and choice modeling that aims to bring machine learning and DCMs closer together (e.g., Sifringer et al. [259], Arkoudi et al. [260], Ramirez et al. [51], van Cranenburgh et al. [261]). We have proposed a new choice model, called "Computer Vision-enriched Discrete Choice Models" (abbreviated as CV-DCM), for modeling multi-attribute choice behavior in the presence of visual and numeric stimuli – methodologically expanding the realm of discrete choice models. The CV-DCM is built from behavioral assumptions, starting with random utility maximization principles. As such, it has a solid behavioral foundation and can be used to derive marginal utilities and (in principle) willingness to pay estimates. The model should thus be conceived as a behavior-informed choice rather than a behavior-agnostic machine learning model. We have demonstrated its merits by applying it to residential location behavior –which is

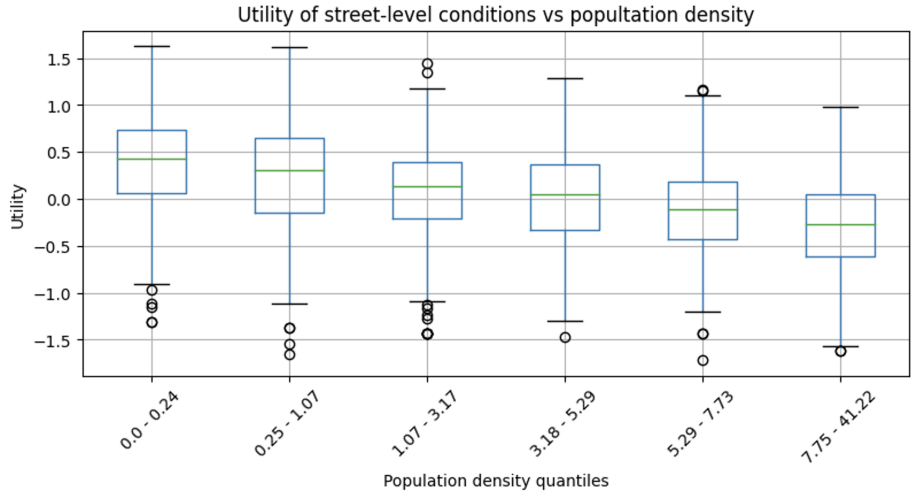



**Figure 7.12** Utility of street-level conditions as a function of population-density quantiles (based on Model 3)

strongly coupled with travel demand. We have shown that CV-DCMs can produce new insights into preferences over visual street-level conditions. Notably, we have uncovered which residential places people find most and least attractive and how attractiveness varies with population density.

7

The proposed model, in conjunction with SC experiments, can potentially enhance the understanding of other transport-related preferences in travel behavior research. Using images can be particularly beneficial when numbers or text fail to convey the choice situation effectively. For instance, preferences related to crowdedness, traffic safety, and spaciousness may be better understood through the use of images in SC experiment showing, e.g. the crowdedness of train platforms, the safety situation of pedestrian crossings, or the extra legroom available when upgrading from economy to business class when booking flights. Incorporating such visualizations can provide valuable insights for transport planners and policymakers seeking to improve transportation systems and services. For instance, inspired by studies like Rossetti et al. [50], in a follow-up study, we used our trained CV-DCM to assess the spatial distribution of utility derived from street-level conditions in residential location choices on a city-wide scale [305].

This study raises a plethora of questions and opens up a multitude of avenues for further investigation at the intersection of choice modeling, computer vision and cognitive psychology. Two technical questions are of particular interest: (1) how to handle multiple images and (2) how to extract more and better information from trained CV-DCMs. Regarding the first question, multiple images are often used to describe alternatives. For instance, real estate websites like Zillow.com and online retailers like Amazon.com often use dozens of images per home or product. While the proposed

modeling framework can accommodate a single image per alternative, future research can extend it to enable multiple images (e.g. inspired by Baevski et al., [306]). This methodological advancement would further expand the application domain and enhance the behavior realism of the discrete choice models. The second question concerns how to extract more and better information from (trained) CV-DCMs. Recent developments in eXplainable AI (XAI) [307] offer a range of techniques that can be adapted to extract information from trained CV-DCMs. In particular, they can be leveraged to shed light on what features are learned by the model to explain the choice behavior and help validate CV-DCMs. Such insights are potentially helpful not only for researchers but also for policymakers and urban planners. For example, in the context of the study's application, XAI techniques should be able to provide insights into the features that make neighborhoods attractive (as shown in Figure 7.8) or unattractive (as shown in Figure 7.9), which can inform the development of planning policies. Additionally, at present, the computation of the standard errors associated with the elements in the feature map (i.e. $\beta_k$) is technically challenging, inter alia, because of large covariance matrices. Future research could explore using smaller feature maps, achieved through techniques like pruning or semantic regularization [308], as potential solutions to address this issue.

We conclude with a word of caution regarding the use of images in choice experiments. Although images hold great potential due to humans' ability to extract information from them effectively, their incorporation into stated choice experiments must be approached cautiously. There are still many uncertainties surrounding their usage. For instance, using images could potentially skew attention to the images (and thus away from the numeric attributes). Its use might thus lead to underestimation of the estimates linked to the numeric attributes. Jansen et al. [309] find some evidence supporting this observation, derived from a (small) survey wherein identical choice tasks were presented with and without accompanying impression photos. In connection with this, there is a risk of biased estimates associated with numeric attributes when CV-DCMs are trained on stated choice data wherein congruence exists between information in the image and numeric attributes (see recent work by Sifringer and Alahi [301]. Hence, care must be taken that the information presented in images does not contain cues about the levels of the numeric attributes when designing stated choice experiments containing images. Another concern regarding the use of images is that people's wishes and preferences may influence their visual perceptions [310]. Simply put, people may see what they want to see. This notion that images can be interpreted in multiple ways is also neatly illustrated by the iconic modern art painting "Ceci n'est pas une pipe". The painting depicts a pipe. However, the artist of the painting, René Magritte, claims that it is not a pipe but a painting (readers interested in a more profound discussion of the painting are referred to Foucault [311]). This highlights the challenge and need to align respondents' interpretation of the images with the researcher's intentions. Keys to the effective use of images in stated choice experiments can likely be found in the cognitive psychology field, which is concerned with studying mental processes such as perception, attention, and memory. Their insights can help researchers in our field to understand better how humans perceive and interpret visual information, which, in turn, can guide, e.g. what

sort of images to use, how to present images (e.g. in relation to numeric attributes), and how to design SC experiments involving images more generally. In sum, further research is needed to comprehensively understand the implications and best practices regarding using images in choice experiments.

Finally, to fully harness the complementary information provided by text and images and pursue the avenues for future research outlined above, it is important to note that our modeling tools need a significant push. The current estimation software, survey platforms, computational resources and data handling practices in our field are not geared towards working with (large numbers of) images. Moreover, working with a large number of images generally places higher technical demands on the programming and data-handling skills of researchers. Fortunately, these hurdles are surmountable. Open science practices and actively seeking cross-fertilization between travel behavior research, choice modeling, computer vision and cognitive psychology can accelerate progress. By sharing our data openly, we hope to contribute to this advancement.

# Chapter 8

# Conclusion

This thesis develops, operationalizes, and demonstrates computer vision as a methodological foundation for an integrated, human-centered urban analysis. In doing so, it bridges the physical structure of cities with their perceptual and behavioral dimensions. Additionally, it clarifies and formalizes what images encode (*components* vs. *conditions*), develops scalable computer vision models to extract and represent urban components, establishes a human-in-the-loop survey infrastructure for capturing perceptual insights, and integrates two fundamental dimensions in urban analytics: what cities *contain* and how they are *perceived*.

These contributions are delivered through six interconnected studies. These studies follow a logical progression from conceptual definition about what images represent, through their computational implementation for capturing physical content and perceptual meaning, to the integration of human perception and experience. One research note (chapter 2) introduces a typology that distinguishes tangible, observable image content from interpretative assessments. Two subsequent studies (chapters 3 and 4) developed computer vision pipelines for extracting and analyzing visible urban components. A dedicated survey-generation platform (chapter 5) was then developed to enable large-scale and reproducible collection of visual perceptual insights and preferences. This platform facilitates the bridge between computational and human approaches. Finally, two concluding studies (chapters 6 and 7) integrate physical and perceptual representations within deep learning, psychological, and behavioral models. Together, these studies position computer vision as a foundational tool in urban analytics.

This chapter elaborates on the specific contributions and conclusions associated with each sub-objective and synthesizes the main findings and conclusions. It concludes with policy recommendations, discusses the limitations, and outlines promising directions for future research.

## 8.1 Specific contributions, findings and conclusions

This section presents the specific contributions, findings, and conclusions associated with each of the four research objectives outlined in Chapter 1. Its purpose is to summarize how each objective was addressed and articulate the main conceptual and methodological insights that emerged from the research.

### 8.1.1 Objective i: Image conceptual clarity

The first objective of this thesis is to clarify and formalize the types of information that images encode. The key is to distinguish what is directly observable from what is perceptually inferred. It also seeks to establish a coherent vocabulary and structure for using imagery in urban analytics, thereby facilitating how future image-based urban research can be scoped and implemented.

This goal is advanced by defining a conceptual typology of the information that images encode. The chapter 2, **Image typology**, differentiates between two primary categories of image content: *components* and *conditions*. Components refer to tangible and measurable visual elements—such as objects, vegetation, buildings, or people. These components can be *explicit*, when they are directly observable in the image, or *implicit*, when they are largely or fully occluded and therefore inferred from spatial relationships or contextual composition. In contrast, conditions denote perceptual, affective, or cognitive qualities inferred by human observers. These can be further divided into *subjective* conditions, which depend on human interpretation, and *objective* conditions, which correspond to measurable environmental states—such as illumination, temperature, or noise level—that can influence their perception. Table 8.1 summarizes this typology in a concise way.

**Table 8.1** Typology of information encoded in images

| Category | Subcategory | Description |
|---|---|---|
| **Components** | *Explicit* | Physical elements that are directly observable in an image and can be pinpointed with pixel coordinates without further assumptions about the context (e.g., buildings, vegetation, vehicles, or people). |
| | *Implicit* | Physical elements that are largely or even fully occluded. They require a bit of reasoning based on contextual cues, arrangement, or composition of explicit components (e.g., density, openness, continuity, or enclosure). |
| **Conditions** | *Subjective* | Perceptual or affective qualities emerging from human interpretation of visual cues (e.g., beauty, order, or vibrancy). They are inherently observer-dependent and culturally influenced. |
| | *Objective* | Environmental or contextual states that influence perception but can be measured independently of human judgment (e.g., illumination intensity in lux, ambient temperature in degrees, or noise level in decibels). |

This typology provides a conceptual and operational foundation for applying computer vision in urban analytics. Specifically, it clarifies which aspects of imagery can be extracted algorithmically and which require human interpretation or perceptual data to be meaningfully represented. As a result, existing and future computer-vision research can be systematically classified within these categories. For example, most object detection studies using street-level imagery [**chen2020**, 32, 34] focus on *explicit components*; some works analyzing occluded objects in images [312] address *implicit components*; the majority of perception-based studies—such as those from the Place Pulse project [22, 47, 48]—examine *subjective conditions* like beauty or vibrancy; while research on lighting perception [313] and other measurable environmental attributes represents *objective conditions*. In practical terms, the latter two categories involve human participants to capture perceptual or experiential information.

Additionally, this typology also establishes a coherent vocabulary for future research, helping to explicitly specify the focus of a study and to clarify what computer-vision models are actually detecting or inferring. By providing such linguistic and conceptual precision, it addresses a key limitation identified in previous research [18], which noted persistent ambiguity and inconsistency in the terminology used to describe visual attributes and perceptual phenomena. This arbitrariness has hindered the comparability and reproducibility of findings across studies. By resolving these conceptual ambiguities, the typology strengthens the theoretical foundation necessary to integrate physical and perceptual dimensions, thereby advancing the overarching goal of this thesis.

### 8.1.2 Objective ii: Computer-vision methods for urban components

The second objective of this thesis is to develop and design computer-vision pipelines focused on capturing urban components. These methods were developed to detect, classify, analyze, and represent the physical elements of the urban environment across large spatial scales. This objective operationalized the first typology (i.e., components) by focusing on the automatic extraction and analysis of the tangible features that structure cities. It specifically addressed how to transform street-level imagery into structured, scalable representations of urban form that can be analyzed and compared consistently across different contexts.



To achieve this goal, two complementary studies are conducted. The first one, *Where Are the People?*, developed a large-scale computer vision pipeline for collecting geo-located street-level imagery, and detecting and counting different urban components in millions of them. Using a pre-trained object detection model, the study automatically extracted instances of human presence and urban-related elements such as vehicles, bicycles, and street furniture. These data were then linked to urban morphological indicators to explore how visible human density varies across and within metropolitan areas. The study thus provided a large-scale, scalable, and image-based pipeline for quantifying urban components, enabling comparative and relationship analysis across them and their spatial configuration.

The second study, *Street Embeddings*, extended this methodological contribution by focusing on the representation and classification of urban form through learned visual embeddings. Using pre-trained convolutional neural networks, the study derived high-dimensional feature vectors for thousands of street-level images, which were then aggregated by street segment. Through clustering, it produced functionally and morphologically coherent street typologies—such as residential, commercial, and arterial corridors—without the need for manual labeling or predefined rules. The results demonstrated that visual embeddings effectively capture patterns of urban structure related to components such as objects, materials, and land-use intensity, revealing the latent organization of the built environment.

Both studies advance the extraction and analysis of urban components from imagery and constitute the methodological core of this thesis with respect to the physical layer of urban environments. They demonstrate that computer-vision models can extract and represent urban components in a scalable and systematic manner. The two approaches, object detection and image embedding, serve complementary purposes. Object detection models allow precise identification and quantification of specific elements within a scene, but are limited in capturing broader compositional relationships. In contrast, embedding models encode the full visual composition of an image into a continuous, high-dimensional representation that captures contextual and relational information, though at the cost of interpretability. Consequently, the selection of models should depend on the analytical goal: object detection is preferable for targeted component analysis, for instance, the inference of pedestrian volumes [33], or the geo-location of traffic signs [34]; whereas embedding models are more suited for capturing holistic urban structure such as inferring urban functional regions [314]. Table 8.2 summarizes the main strengths and limitations of these two ways of capturing components.

### 8.1.3 Objective iii: Human-in-the-loop measurement of conditions



The third objective of this thesis is to develop a platform that facilitates the creation and deployment of image-based surveys. This addresses the collection of the perceptual dimension of the typology. While computer vision can approximate perceptual attributes, human judgments are essential for capturing subjective qualities such as safety, beauty, or comfort. Therefore, this objective focused on enabling scalable, reproducible, and ethically sound methods for eliciting perceptual data linked to imagery.

To achieve this, *PixelSurvey* is developed. This is a modular, open-source platform that supports the creation and implementation of image-based surveys with human-input experiments. The software allows researchers to create tasks such as stated choice, similarity judgments, and Likert-scale experiments using street-level imagery. PixelSurvey automates web development, experiment implementation, and response storage. In this way, the platform standardizes workflows, enhances reproducibility, and ensures transparency across perception-based studies.

**Table 8.2** Comparison of computer-vision approaches for extracting urban components from imagery

| Approach | Strengths | Limitations |
|---|---|---|
| **Object detection models** | Provide explicit identification and localization of specific components. Enable direct quantification and spatial analysis of visible elements with clear interpretability. Well-suited for targeted urban indicators and mapping applications. | Depend on the availability and quality of annotated training data. Performance varies across contexts, camera angles, and illumination conditions. Capture individual elements but not their broader spatial or compositional relationships. |
| **Image embedding models** | Generate high-dimensional representations of entire scenes that capture contextual and relational information among components. Allow clustering, similarity search, and typology generation without manual labeling. Transferable across cities and scalable to large datasets. | Difficult to interpret, as latent features are abstract and not directly linked to human-understandable categories. Do not provide explicit localization of objects. Require careful aggregation and validation to ensure semantic consistency across contexts. |

PixelSurvey facilitates the integration of imagery into survey-based research and opens new avenues for developing domain-specific computer-vision applications. By lowering the technical barriers to conducting image-based experiments, the platform enables disciplines such as discrete choice modeling [156, 157] to incorporate visual stimuli as part of their experimental design, advancing the development of perception-aware choice models. Furthermore, PixelSurvey provides a bridge between behavioral science and state-of-the-art computer vision models [2, 25, 179]. This framework allows the collection of human input from images to train or adapt these advanced models across domains. Similar approaches have already shown value in other fields where imagery is a central medium of analysis, such as landscape perception [263], or cognitive psychology [175], suggesting that PixelSurvey can serve as an enabler for cross-disciplinary research integrating visual data and human experience.

### 8.1.4 Objective iv: Integration of components and conditions for urban understanding

The fourth objective of this thesis is to integrate information on components and conditions within an analytical framework. In doing so, it develops models capable of extracting, structuring, and predicting perceptions and preferences based on the overall visual composition of urban imagery and human inputs. This objective therefore, addresses the central ambition of the thesis: to connect what cities *contain* with how they are *perceived*, and to operationalize this connection for analytical purposes. Concretely, image-embedding models produce vector representations that encode the visual composition of street-level scenes, mapping what cities contain, while human-in-the-loop tasks (e.g., similarity judgments and choice experiments) provide perceptual supervision and preference data that let the models learn how those places are perceived.

Two studies are conducted to achieve this goal. The first one, *From Pixels to Perceptions*, used human similarity judgments to guide urban representation learning. Participants compared triplets of urban scenes, identifying the image that is most dissimilar from the other two. These human-provided similarity relationships were then used to supervise the training of an embedding model. The resulting embedding space aligns machine-learned representations with perceptual structure. The results show that models trained with perceptual supervision capture finer nuances in how people group and differentiate urban environments compared to computer vision models without human inputs. Additionally, physical components and perceptual conditions are integrated within those representations.

The second study, *Computer Vision–Enriched Discrete Choice Models (CV–DCM)*, extended this integration toward behavioral analysis. In this case, participants made residential location choices between alternatives described by both urban images and numerical attributes. Image embeddings extracted from street-level imagery were combined with discrete choice models based on Random Utility Maximization. This allows the embedding model to jointly learn visual feature representations (i.e., components) and human preferences (based on components and conditions). This approach enables the interpretation of the relevant numerical attributes in relation to choices, taking into account visual information encoded in images.



Both studies demonstrate the feasibility and value of coupling visual, perceptual, and behavioral information within an the same pipeline. From Pixels to Perceptions contributes to a growing body of research that seeks to align computational representations with human cognition [175, 315]. It extends this paradigm to urban analysis by embedding perceptual similarity directly into image representations of cities. By doing so, it produces embeddings with perception-aligned visual components and conditions that reflect how people intuitively group and interpret the built environment and their characteristics. Meanwhile, CV–DCM expands the methodological frontier of discrete choice modeling [157, 188] by incorporating computer-vision-derived features as perceptual descriptors

of urban alternatives. This integration of visual embeddings with Random Utility Maximization enables behavioral models to account for visual and experiential cues that are otherwise unobservable in conventional numerical datasets.

## 8.2 Pixels · People · Places: Overarching conclusions

This thesis has developed, operationalized, and demonstrated the use of computer vision as a methodological foundation for integrated and human-centered urban analytics. It bridges the physical structure of cities with their perceptual and behavioral dimensions, and establishes an integrated framework for understanding how urban form is both seen and perceived. Additionally, it has positioned Artificial intelligence (AI) as a relevant tool in urban studies for understanding how visual form shapes perception, preference, and behavior. Importantly, emphasizing that such understanding is only possible through the active participation of people in the loop—whose perceptions, judgments, and choices ground computational analysis in human interpretation.

Images and computer vision methods have the potential to improve urban analytics. First, images can reveal aspects of cities that traditional data often overlook. Through image-based experiments, people can freely reflect and observe dimensions that are not initially considered in studies. Therefore, images open the possibility of discovering latent or emergent dimensions of urban quality that are not initially anticipated by the analyst. In this sense, visual data act as generative evidence, inviting interpretation, contextualization, and refinement through human judgment. Additionally, unlike tabular indicators, street-level imagery shows places from a human viewpoint and retains the cues people based their perceptions (e.g., enclosure, greenery, facade quality). This provides a common reference for researchers and residents to compare, rate, and discuss urban environments. This visual proximity allows urban analysis to integrate the physical structure of cities with the perceptual and behavioral responses they evoke.

The thesis also demonstrates that human participation is essential for transforming visual data into meaningful urban knowledge. Computer vision can detect what is physically observable, but only human-in-the-loop approaches can capture what is felt, evaluated, or preferred in specific contexts—and these perceptions vary across places, cultures, and populations. Incorporating human interpretation ensures that models remain socially grounded and locally relevant, bridging computational precision with lived urban diversity.



Ultimately, the thesis advances the view that cities must be analyzed both visually and perceptually. Computer vision becomes not just a technological instrument but a methodological bridge that connects **pixels** with **people** to understand **places**. By positioning imagery at the intersection of observation and interpretation, this research establishes the conceptual foundation for perception-aware urban analytics: a field that seeks to understand cities through the intertwined lenses of form, meaning, and human response.

In sum, this thesis shows how computer vision, when coupled with human judgment, turns pixels into policy-relevant evidence. By systematically connecting pixels, people, and places, it lays the groundwork for perception-aware urban analytics that is both scientifically rigorous and socially responsive.

## 8.3 Policy recommendations

The findings of this thesis demonstrate that computer vision, when grounded in human perception, can support scalable, socially relevant urban analytics. Yet deploying these methods in policy and planning contexts requires caution because of concerns about interpretability, inclusivity, and ethical implications of visual data. In light of these challenges, this section offers concrete recommendations to guide the responsible translation of perception-aware urban analytics into public decision-making.

First, urban indicators derived from street-level imagery (e.g., people's density, physical elements, or perceptual qualities) can support evidence-based planning if properly contextualized. These indicators are spatially explicit and scalable, enabling continuous monitoring of urban vitality, aesthetic quality, and livability. Planners and municipalities can integrate these indicators to inform interventions, prioritize investment in public spaces, and evaluate spatial inequalities across neighborhoods. For example, the integration of image-derived measures of human presence, as demonstrated in chapter 3: *Where Are the People?*, the assessment of urban activity patterns beyond traditional mobility datasets. This type of visual evidence can complement conventional indicators and offer a richer picture of how people interact with the built environment.

Second, tools like PixelSurvey enable a more inclusive form of citizen engagement by allowing people to express how they perceive their environments through images, rather than technical or verbal means. This lowers participation barriers and enhances representativeness in participatory planning processes. In contexts where technical language or abstract planning concepts create exclusion, image-based platforms provide a shared visual language for dialogue. Policymakers should adopt such tools to democratize the planning process, enabling communities to actively co-shape interventions in their neighborhoods.

Third, policymakers and researchers should ensure geographic and demographic representativeness in image data and adopt fairness-aware sampling and evaluation before using image-based perception models in policy evaluation. Most publicly available imagery is concentrated in specific regions, particularly in affluent or well-documented areas, leading to biased model generalization. AI systems trained on these datasets may fail to represent informal settlements or under-mapped areas, potentially reinforcing existing spatial and social inequities. This bias is not merely technical—it has practical implications when such models inform decisions about urban renewal, transport planning, or public safety. Policymakers and researchers must ensure geographic and demographic representativeness in image data and incorporate fairness-aware sampling strategies.

Fourth, image data in public sector should adopt enforceable protocols for privacy, but it should be shared for academic and research use when the purpose is clear and justified. Street-level imagery often contains identifiable individuals, private residences, or sensitive locations. On the other hand, oftentimes platforms and institutions own the images. These issues create barriers for collaboration, future research and policy usage. It is imperative to anonymize data, and the use of such images in research and policy must comply with data protection regulations and adhere to principles of consent and transparency. Agencies and universities should adopt simple, written protocols that require de-identification (blur faces, license plates, house numbers), limit who can access original images and for how long, prefer releasing summaries or embeddings instead of raw photos, and use data-sharing agreements that specify purpose, security, retention, and licensing. Following privacy law, consent, and transparency requirements enables collaboration and policy use without compromising individual rights.

Finally, the use perception-aware analytics as a standard tool for ex ante appraisal of the experiential impacts of proposed urban interventions. By linking visual indicators with interpretable behavioral models, as demonstrated in the *CV–DCM* study, policymakers and agencies can estimate how changes in the built environment may shift perceptions, preferences, and choices. This brings the perceptual dimension into formal appraisal and helps infer its social benefits: by tying perceptions to specific features within choice models, planners can estimate effect sizes and translate them into willingness-to-pay for social cost–benefit analysis. This enables ex ante evaluation of both functional and experiential impacts for streetscape redesigns, greening initiatives, and public-space investments, helping close the gap between design intentions and lived experience.

## 8.4 Limitations and future research directions

This thesis proposes an integrated, perception-aware approach to urban analytics, yet several limitations are present in the methodological and conceptual contributions of this work. These limitations offer promising directions for future research in urban analytics. This subsection discusses the main limitations alongside associated future research directions.



A first limitation concerns the boundary between *components* and *conditions*. The typology distinguishes observable content from inferred qualities, yet boundary cases (occlusion, context dependence, cultural variation) can blur this distinction and yield ambiguous classifications. For instance, a *chair* which is a component based on our typology, can also be a *table*. So, even components can be affected by interpretation. This difference in meaning could produce inconsistencies in some studies. Therefore, it is important to define clearly the concepts before investigating with them. Future research should explore how to operationalize the typology further. For example by defining rules or exploring the possible cultural effects on those different interpretations.

A second limitation involves the coverage, aggregation, and generalization biases inherent to imagery-based workflows. Specifically, those arising from coverage, aggregation, and generalization in imagery-based pipelines. Street-level imagery is

uneven in space, time, and quality, and domain shift across seasons, sensors, view angles, and morphologies degrades detection and embedding performance. This can affect the quality and generalization of the information generated from images. As a result, transferability across cities—and fairness across under-imaged areas—remains imperfect. Future work should broaden sources (e.g., multi-platform SLI), incorporate time-series imagery to capture diurnal and seasonal dynamics, and pair visual data with complementary data modalities (POIs, mobility, remote sensing) within domain-adaptive and uncertainty-aware models. Systematic sensitivity analyses across spatial units and sampling strategies, coupled with active sampling in data-scarce zones, can reduce single-source dependency and strengthen generalization.

A third limitation involves interpretability in representation learning and its behavioral integration. Image embeddings themselves capture components, conditions, and their context but those are opaque; when used within behavioral models, latent vector weights lack direct policy meaning, limiting interpretability and design guidance. Future research could develop hybrid models that pair interpretable object- and scene-level features with embeddings, apply ante-hoc and post-hoc XAI (e.g., concept bottlenecks, feature attribution, counterfactual visuals), and use human-in-the-loop audits to map latent directions to actionable urban attributes (e.g., enclosure, greenery share, traffic presence). Reporting standards—model cards, saliency summaries, and stability checks—should accompany perception-aware models to support scrutiny and reuse.

A fourth limitation concerns the gap between measured perceptions and lived experience. Perceptual indicators and image-based judgments are informative, yet they are inevitably proxies. Subjective experience contains aspects that elude complete third-person description. By analogy, no finite set of CV features, embeddings, or survey responses can fully capture the phenomenology of being in a place [316]. Moreover, perceptions are highly heterogeneous across individuals and temporally variable within the same individual (mood, purpose of trip, companions, weather, life stage) [317]. Current methods, even with rich embeddings and choice models, only partially recover this latent diversity. Future work should combine population-level scalability with within-person longitudinal designs and richer context situations (e.g., purpose of the perception).



Finally, ethics, equity, and governance apply to all parts of this work. Privacy, consent, and licensing rules limit how images can be shared. Coverage is uneven across places and times, which can reproduce inequities. Large AI pipelines also have computing and environmental costs. To address these issues, it is imperative to build governance into the workflow. For instance, publish a short dataset statement for every dataset, explore and report bias and coverage about datasets, involve local communities to review outputs and record any changes, and release reproducible code, environment files, and model/data cards. These steps make clear what can be studied, where, and for whom, and they help build trust in perception-aware urban analytics.

Together, these limitations highlight that while the proposed methods significantly expand the analytical reach of urban research, their responsible application requires careful consideration of interpretability, equity, and ethics. Acknowledging these

boundaries can be a relevant opportunity as it defines a road map for improving transparency, cross-cultural validity, and policy relevance in future computer-vision-based urban studies.

## 8.5 My concluding remarks

We are living through a big wave of AI. For urban policy, this brings real chances to improve how we observe cities, build stronger evidence, and test ideas *ex ante*. However, it also brings many risks: confusing prediction with understanding, relying on black-box systems, and repeating existing inequalities. A simple rule helps: start from problems, not tools; use evidence before scaling; keep people before pixels.

As researchers, our responsibilities do not change with each new model. We should be clear about what images can show—and honest about what they cannot. We should measure uncertainty, check coverage and bias, protect privacy, and ask communities to review results about their places. Sharing code, documentation, and knowledge.

Cities are local. The same visual feature can mean different things in different neighborhoods or cultures. We should not assume models transfer; we should test and adapt them. Models should also explain which features matter and why, so design and policy trade-offs are visible and open to debate.

AI will keep evolving, and the specific algorithms we use today will inevitably become obsolete. However, our principles and analytical frameworks must remain atemporal: transparency, accountability, inclusion, respect for local context, and the fundamental distinction between what a machine observes and what a human feels. By anchoring computational tools in these enduring concepts, this thesis provides a future-proof method. It ensures that as technology advances, computer vision will continue to connect *pixels, people, and places*—informing urban decisions and empowering, rather than replacing, human judgment. This thesis is one step in that direction.

# Bibliography


[1] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan. 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions'. In: *Journal of big Data* 8.1 (2021), p. 53.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al. 'An image is worth 16x16 words: Transformers for image recognition at scale'. In: *arXiv preprint arXiv:2010.11929* (2020).

[3] T. Mikolov. 'Efficient estimation of word representations in vector space'. In: *arXiv preprint arXiv:1301.3781* (2013).

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al. 'Language models are few-shot learners'. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. 'Generative adversarial nets'. In: *Advances in neural information processing systems* 27 (2014).

[6] Y. LeCun, Y. Bengio and G. Hinton. 'Deep learning'. In: *nature* 521.7553 (2015), pp. 436–444.

[7] M. R. Ibrahim, J. Haworth and T. Cheng. 'Understanding cities with machine eyes: A review of deep computer vision in urban analytics'. In: *Cities* 96 (2020), p. 102481.

[8] A. Paivio. 'Mental imagery in associative learning and memory.' In: *Psychological review* 76.3 (1969), p. 241.

[9] Google. *Google Street View*. Accessed: 2024-11-03. 2023. URL: `https://www.google.com/maps/streetview/`.

[10] Apple Inc. *Apple Maps Look Around*. Accessed: 2024-04-12. 2023. URL: `https://www.apple.com/maps/`.

[11] Tencent. *Tencent Maps*. `https://map.qq.com/`. Accessed: 2022-09-10. 2022.

[12] S. Ren, K. He, R. Girshick and J. Sun. 'Faster r-cnn: Towards real-time object detection with region proposal networks'. In: *Advances in neural information processing systems* 28 (2015).

[13] J. Long, E. Shelhamer and T. Darrell. 'Fully convolutional networks for semantic segmentation'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[14] G. Wu, W. Lu, G. Gao, C. Zhao and J. Liu. 'Regional deep learning model for visual tracking'. In: *Neurocomputing* 175 (2016), pp. 310–323.

[15] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba. 'Places: A 10 million image database for scene recognition'. In: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), pp. 1452–1464.

[16] A. Krizhevsky, I. Sutskever and G. E. Hinton. 'Imagenet classification with deep convolutional neural networks'. In: *Advances in neural information processing systems* 25 (2012).

[17] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen and Y. Wu. 'Learning fine-grained image similarity with deep ranking'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1386–1393.

[18] L. Liu and A. Sevtsuk. 'Clarity or confusion: A review of computer vision street attributes in urban studies and planning'. In: *Cities* 150 (2024), p. 105022.

[19] M. De Nadai, R. L. Vieriu, G. Zen, S. Dragicevic, N. Naik, M. Caraviello, C. A. Hidalgo, N. Sebe and B. Lepri. 'Are safer looking neighborhoods more lively? A multimodal investigation into urban life'. In: *Proceedings of the 24th ACM international conference on Multimedia*. 2016, pp. 1127–1135.

[20] B.-S. Kweon, J. Rosenblatt-Naderi, C. D. Ellis, W.-H. Shin and B. H. Danies. 'The effects of pedestrian environments on walking behaviors and perception of pedestrian safety'. In: *Sustainability* 13.16 (2021), p. 8728.

[21] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent and J. Weaver. 'Google street view: Capturing the world at street level'. In: *Computer* 43.6 (2010), pp. 32–38.

[22] A. Dubey, N. Naik, D. Parikh, R. Raskar and C. A. Hidalgo. 'Deep learning the city: Quantifying urban perception at a global scale'. In: *European conference on computer vision*. Springer. 2016, pp. 196–212.

[23] D. Quercia, R. Schifanella and L. M. Aiello. 'The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city'. In: *Proceedings of the 25th ACM conference on Hypertext and social media*. 2014, pp. 116–125.

[24] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner. 'Gradient-based learning applied to document recognition'. In: *Proceedings of the IEEE* 86.11 (2002), pp. 2278–2324.

[25] K. He, X. Zhang, S. Ren and J. Sun. 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[26] M. D. Zeiler and R. Fergus. 'Visualizing and understanding convolutional networks'. In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[27] J. Redmon, S. Divvala, R. Girshick and A. Farhadi. 'You only look once: Unified, real-time object detection'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[28] G. E. Hinton and R. R. Salakhutdinov. 'Reducing the dimensionality of data with neural networks'. In: *science* 313.5786 (2006), pp. 504–507.

[29] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol and L. Bottou. 'Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.' In: *Journal of machine learning research* 11.12 (2010).

[30] Y. Taigman, M. Yang, M. Ranzato and L. Wolf. 'Deepface: Closing the gap to human-level performance in face verification'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1701–1708.

[31] F. Schroff, D. Kalenichenko and J. Philbin. 'Facenet: A unified embedding for face recognition and clustering'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[32] K. Choi, W. Lim, B. Chang, J. Jeong, I. Kim, C.-R. Park and D. W. Ko. 'An automatic approach for tree species detection and profile estimation of urban street trees using deep learning and Google street view images'. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 190 (2022), pp. 165–180.

[33] L. Chen, Y. Lu, Q. Sheng, Y. Ye, R. Wang and Y. Liu. 'Estimating pedestrian volume using Street View images: A large-scale validation test'. In: *Computers, Environment and Urban Systems* 81 (2020), p. 101481.

[34] A. Campbell, A. Both and Q. C. Sun. 'Detecting and mapping traffic signs from Google Street View images using deep learning and GIS'. In: *Computers, Environment and Urban Systems* 77 (2019), p. 101350.

[35] S. Qiu, A. Psyllidis, A. Bozzon and G.-J. Houben. 'Crowd-mapping urban objects from street-level imagery'. In: *The world wide web conference*. 2019, pp. 1521–1531.

[36] G. Dong, Y. Yan, C. Shen and H. Wang. 'Real-time high-performance semantic image segmentation of urban street scenes'. In: *IEEE Transactions on Intelligent Transportation Systems* 22.6 (2020), pp. 3258–3274.

[37] A. Chaurasia and E. Culurciello. 'Linknet: Exploiting encoder representations for efficient semantic segmentation'. In: *2017 IEEE visual communications and image processing (VCIP)*. IEEE. 2017, pp. 1–4.

[38] A. Bottino, A. Garbo, C. Loiacono and S. Quer. 'Street viewer: An autonomous vision based traffic tracking system'. In: *Sensors* 16.6 (2016), p. 813.

[39] X. Li, C. Zhang, W. Li, R. Ricard, Q. Meng and W. Zhang. 'Assessing street-level urban greenery using Google Street View and a modified green view index'. In: *Urban forestry & urban greening* 14.3 (2015), pp. 675–685.

[40] Y. Xia, N. Yabuki and T. Fukuda. 'Development of a system for assessing the quality of urban street-level greenery using street view images and deep learning'. In: *Urban Forestry & Urban Greening* 59 (2021), p. 126995.

[41] M. B. Starzyńska-Grześ, R. Roussel, S. Jacoby and A. Asadipour. 'Computer vision-based analysis of buildings and built environments: A systematic review of current approaches'. In: *ACM Computing Surveys* 55.13s (2023), pp. 1–25.

[42] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser and C. A. Hidalgo. 'Computer vision uncovers predictors of physical urban change'. In: *Proceedings of the National Academy of Sciences* 114.29 (2017), pp. 7571–7576.

[43] Z. Wang, H. Li and R. Rajagopal. 'Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 1013–1020.

[44] Y. Luo, C.-T. Leong, S. Jiao, F.-L. Chung, W. Li and G. Liu. 'Geo-Tile2Vec: A multi-modal and multi-stage embedding framework for urban analytics'. In: *ACM Transactions on Spatial Algorithms and Systems* 9.2 (2023), pp. 1–25.

[45] T. Huang, Z. Wang, H. Sheng, A. Y. Ng and R. Rajagopal. 'M3G: Learning urban neighborhood representation from multi-modal multi-graph'. In: *Proceedings of the DeepSpatial* 2021 (2021).

[46] S. Woźniak and P. Szymański. 'Hex2vec: Context-aware embedding h3 hexagons with openstreetmap tags'. In: *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. 2021, pp. 61–71.

[47] N. Naik, J. Philipoom, R. Raskar and C. Hidalgo. 'Streetscore-predicting the perceived safety of one million streetscapes'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 779–785.

[48] M. Yu, X. Chen, X. Zheng, W. Cui, Q. Ji and H. Xing. 'Evaluation of spatial visual perception of streets based on deep learning and spatial syntax'. In: *Scientific Reports* 15.1 (2025), p. 18439.

[49] T. Van Asten, V. Milias, A. Bozzon and A. Psyllidis. '"Eyes on the Street": Estimating Natural Surveillance Along Amsterdam's City Streets Using Street-Level Imagery'. In: *International Conference on Computers in Urban Planning and Urban Management*. Springer. 2023, pp. 215–229.

[50] T. Rossetti, H. Lobel, V. Rocco and R. Hurtubia. 'Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach'. In: *Landscape and urban planning* 181 (2019), pp. 169–178.

[51] T. Ramírez, R. Hurtubia, H. Lobel and T. Rossetti. 'Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety'. In: *Landscape and Urban Planning* 208 (2021), p. 104002.

[52] R. Marasinghe, T. Yigitcanlar, S. Mayere, T. Washington and M. Limb. 'Computer vision applications for urban planning: A systematic review of opportunities and constraints'. In: *Sustainable Cities and Society* 100 (2024), p. 105047.

[53] P. Florio, T. Leduc, Y. Sutter and R. Brémond. 'Visual complexity of urban streetscapes: human vs computer vision'. In: *Machine Vision and Applications* 35.1 (2024), p. 7.

[54] A. Vanky and R. Le. 'Urban-semantic computer vision: a framework for contextual understanding of people in urban spaces'. In: *AI & SOCIETY* 38.3 (2023), pp. 1193–1207.

[55] X. Fu, T. Jia, X. Zhang, S. Li and Y. Zhang. 'Do street-level scene perceptions affect housing prices in Chinese megacities? An analysis using open access datasets and deep learning'. In: *PloS one* 14.5 (2019), e0217505.

[56] R. H. Matsuoka and R. Kaplan. 'People needs in the urban landscape: analysis of landscape and urban planning contributions'. In: *Landscape and urban planning* 84.1 (2008), pp. 7–19.

[57] W. Min, S. Mei, L. Liu, Y. Wang and S. Jiang. 'Multi-task deep relative attribute learning for visual urban perception'. In: *IEEE Transactions on Image Processing* 29 (2019), pp. 657–669.

[58] K. Lynch. *The image of the city*. MIT press, 1964.

[59] J. Gehl. *Life between buildings.* Danish Architectural Press, 2011.

[60] A. Aiello, R. G. Ardone and M. Scopelliti. 'Neighbourhood planning improvement: Physical attributes, cognitive and affective evaluation and activities in two neighbourhoods in Rome'. In: *Evaluation and Program Planning* 33.3 (2010), pp. 264–275.

[61] M. Bonaiuto, A. Aiello, M. Perugini, M. Bonnes and A. P. Ercolani. 'Multidimensional perception of residential environment quality and neighbourhood attachment in the urban environment'. In: *Journal of environmental psychology* 19.4 (1999), pp. 331–352.

[62] R. García-Mira, C. Arce and J. M. Sabucedo. 'Perceived quality of neighbourhoods in a city in northwest Spain: an individual differences scaling approach'. In: *Journal of environmental psychology* 17.3 (1997), pp. 243–252.

[63] F. Biljecki and K. Ito. 'Street view imagery in urban analytics and GIS: A review'. In: *Landscape and Urban Planning* 215 (2021), p. 104217.

[64] Y. Lu. 'Using Google Street View to investigate the association between street greenery and physical activity'. In: *Landscape and Urban Planning* 191 (2019), p. 103435.

[65] J. Tang and Y. Long. 'Measuring visual quality of street space and its temporal variation: Methodology and its application in the Hutong area in Beijing'. In: *Landscape and Urban Planning* 191 (2019), p. 103436.

[66] X. Gu, W. Xu, C. Gong and X. Liu. 'City centers really lived up to the hype? Evidence from human perceptions of over 4000 communities in China'. In: *Cities* 166 (2025), p. 106278.

[67] L. Dupont, M. Antrop and V. Van Eetvelde. 'Does landscape related expertise influence the visual perception of landscape photographs? Implications for participatory landscape planning and management'. In: *Landscape and Urban Planning* 141 (2015), pp. 68–77.

[68] M. Franěk, J. Petružálek and D. Šefara. 'Eye movements in viewing urban images and natural images in diverse vegetation periods'. In: *Urban Forestry & Urban Greening* 46 (2019), p. 126477.

[69] J. B. Hollander, A. Purdy, A. Wiley, V. Foster, R. J. Jacob, H. A. Taylor and T. T. Brunyé. 'Seeing the city: Using eye-tracking technology to explore cognitive responses to the built environment'. In: *Journal of Urbanism: International Research on Placemaking and Urban Sustainability* (2019).

[70] P. Kay and T. Regier. 'Language, thought and color: recent developments'. In: *Trends in cognitive sciences* 10.2 (2006), pp. 51–54.

[71] P. Cunha and D. C. Moura. 'A scalable and privacy preserving approach for counting pedestrians in urban environment'. In: *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2015, pp. 1–6.

[72] R. Krier and C. Rowe. *Urban space*. Academy editions London, 1979.

[73] L. Lebel, C. Krittasudthacheewa, A. Salamanca and P. Sriyasak. 'Lifestyles and consumption in cities and the links with health and well-being: the case of obesity'. In: *Current opinion in environmental sustainability* 4.4 (2012), pp. 405–413.

[74] Q. Sheng, D. Wan and B. Yu. 'Effect of Space Configurational Attributes on Social Interactions in Urban Parks'. In: *Sustainability* 13.14 (2021), p. 7805.

[75] Z. I. Abass and R. Tucker. 'Talk on the street: the impact of good streetscape design on neighbourhood experience in low-density suburbs'. In: *Housing, Theory and Society* 38.2 (2021), pp. 204–227.

[76] Z. I. Abass and R. Tucker. 'Fifty shades of green: tree coverage and neighbourhood attachment in relation to social interaction in Australian suburbs'. In: *ASA 2016: Revisiting the Role of Architectural Science in Design and Practice: Proceedings of the 50th International Conference of the Architectural Science Association*. University of Adelaide. 2016, pp. 259–268.

[77] K. Krellenberg, J. Welz and S. Reyes-Päcke. 'Urban green areas and their potential for social interaction–A case study of a socio-economically mixed neighbourhood in Santiago de Chile'. In: *Habitat International* 44 (2014), pp. 11–21.

[78] A. Uslu et al. 'Social interaction in urban transformation areas and the characteristics of urban outdoor spaces: a case study from Turkey'. In: *African Journal of Agricultural Research* 5.20 (2010), pp. 2801–2810.

[79] K. Mouratidis. 'Built environment and social well-being: How does urban form affect social life and personal relationships?' In: *Cities* 74 (2018), pp. 7–20.

[80] T. Sugiyama and C. W. Thompson. 'Environmental support for outdoor activities and older people's quality of life'. In: *Journal of Housing for the Elderly* 19.3-4 (2006), pp. 167–185.

[81] B. Lipovská et al. 'Assessing observation methods for landscape planning practice in rural villages'. In: *Current Urban Studies* 1.04 (2013), p. 102.

[82] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga et al. 'Social contacts and mixing patterns relevant to the spread of infectious diseases'. In: *PLoS medicine* 5.3 (2008), e74.

[83] Z. A. Hamstead, D. Fisher, R. T. Ilieva, S. A. Wood, T. McPhearson and P. Kremer. 'Geolocated social media as a rapid indicator of park visitation and equitable park access'. In: *Computers, Environment and Urban Systems* 72 (2018), pp. 38–50.

[84] G. McKenzie, K. Janowicz, S. Gao and L. Gong. 'How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest'. In: *Computers, Environment and Urban Systems* 54 (2015), pp. 336–346.

[85] E. Steiger, R. Westerholt, B. Resch and A. Zipf. 'Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data'. In: *Computers, environment and urban systems* 54 (2015), pp. 255–265.

[86] S. Bocconi, A. Bozzon, A. Psyllidis, C. Titos Bolivar and G.-J. Houben. 'Social glass: A platform for urban analytics and decision-making through heterogeneous social data'. In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, pp. 175–178.

[87] M. W. Traunmueller, N. Johnson, A. Malik and C. E. Kontokosta. 'Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities'. In: *Computers, Environment and Urban Systems* 72 (2018), pp. 4–12.

[88] C. E. Kontokosta and N. Johnson. 'Urban phenology: Toward a real-time census of the city using Wi-Fi data'. In: *Computers, Environment and Urban Systems* 64 (2017), pp. 144–153.

[89] P. Danielis, S. T. Kouyoumdjieva and G. Karlsson. 'Urbancount: Mobile crowd counting in urban environments'. In: *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE. 2017, pp. 640–648.

[90] M. B. Shami, S. Maqbool, H. Sajid, Y. Ayaz and S.-C. S. Cheung. 'People counting in dense crowd images using sparse head detections'. In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.9 (2018), pp. 2627–2636.

[91] M. Jendryke, T. Balz, S. C. McClure and M. Liao. 'Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai'. In: *Computers, Environment and Urban Systems* 62 (2017), pp. 99–112.

[92] A. Bansal and K. Venkatesh. 'People counting in high density crowds from still images'. In: *arXiv preprint arXiv:1507.08445* (2015).

[93] Mapillary. *Mapillary*. `https://www.mapillary.com/`. Accessed: 2022-09-10. 2022.

[94] OSM. *Planet dump retrieved from https://osm.org*. `https://www.openstreetmap.org`. 2022.

[95] G. Boeing. 'OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks'. In: *Computers, Environment and Urban Systems* 65 (2017), pp. 126–139.

[96] OSM Wiki. *Map features — OpenStreetMap Wiki*. [Online; accessed 15-October-2022]. 2022. URL: `https://wiki.openstreetmap.org/w/index.php?title=Map_features&oldid=2420438`.

[97] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan et al. 'Searching for mobilenetv3'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1314–1324.

[98] J. Durbin. 'Estimation of parameters in time-series regression models'. In: *Journal of the royal statistical society: Series B (Methodological)* 22.1 (1960), pp. 139–153.

[99] W. Tobler. 'Philosophy in geography'. In: *Theory and Decision Library* 20.1 (1979), pp. 379–386.

[100] CBS. *Population*. `https://www.cbs.nl/en-gb/figures/detail/37296eng`. Accessed: 2022-09-10. 2022.

[101] S. Openshaw and P. Taylor. 'The modified areal unit problem'. In: *Quantitative Geography* (1984), pp. 60–69.

[102] G. Arbia and F. Petrarca. 'Effects of MAUP on spatial econometric models'. In: *Letters in Spatial and Resource Sciences* 4 (2011), pp. 173–185.

[103] J. Jacobs. *The Death and Life of Great American Cities*. Vintage Books ed. Knopf Doubleday Publishing Group, 1961. URL: https://books.google.nl/books?id=P%5C_bPTgOoBYkC.

[104] I. Gómez-Varo, X. Delclòs-Alió and C. Miralles-Guasch. 'Jane Jacobs reloaded: A contemporary operationalization of urban vitality in a district in Barcelona'. In: *Cities* 123 (2022), p. 103565.

[105] T. Hägerstrand. *What about people in regional science, regional science association papers, Vol. XXIV*. 1970.

[106] A. Zhang, W. Li, J. Wu, J. Lin, J. Chu and C. Xia. 'How can the urban landscape affect urban vitality at the street block level? A case study of 15 metropolises in China'. In: *Environment and Planning B: Urban Analytics and City Science* 48.5 (2021), pp. 1245–1262.

[107] R. Askarizad and H. Safari. 'The influence of social interactions on the behavioral patterns of the people in urban spaces (case study: The pedestrian zone of Rasht Municipality Square, Iran)'. In: *Cities* 101 (2020), p. 102687.

[108] CycloMedia. *Cyclomedia Home Page*. https://www.cyclomedia.com. Accessed: 2022-09-10. 2022.

[109] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia. 'Pyramid Scene Parsing Network'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[110] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko. 'End-to-end object detection with transformers'. In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[111] A. Getis and J. Getis. 'Christaller's central place theory'. In: *Journal of Geography* 65.5 (1966), pp. 220–226.

[112] G. K. Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[113] R. P. Lopez and H. P. Hynes. 'Obesity, physical activity, and the urban environment: public health research needs'. In: *Environmental Health* 5.1 (2006), pp. 1–10.

[114] J. F. Sallis, E. Cerin, T. L. Conway, M. A. Adams, L. D. Frank, M. Pratt, D. Salvo, J. Schipperijn, G. Smith, K. L. Cain et al. 'Physical activity in relation to urban environments in 14 cities worldwide: a cross-sectional study'. In: *The lancet* 387.10034 (2016), pp. 2207–2217.

[115] A. Birenboim, M. Helbich and M.-P. Kwan. 'Advances in portable sensing for urban environments: Understanding cities from a mobility perspective'. In: *Computers, Environment and Urban Systems* 88 (2021), p. 101650.

[116] W. Lee, H. Kim, H. M. Choi, S. Heo, K. C. Fong, J. Yang, C. Park, H. Kim and M. L. Bell. 'Urban environments and COVID-19 in three Eastern states of the United States'. In: *Science of The Total Environment* 779 (2021), p. 146334.

[117] A. D. Singleton and P. A. Longley. 'Data infrastructure requirements for new geodemographic classifications: The example of London's workplace zones'. In: *Applied Geography* 109 (2019), p. 102038.

[118] H. Christian, A. Bauman, J. N. Epping, G. N. Levine, G. McCormack, R. E. Rhodes, E. Richards, M. Rock and C. Westgarth. 'Encouraging dog walking for health promotion and disease prevention'. In: *American journal of lifestyle medicine* 12.3 (2018), pp. 233–243.

[119] I. Chatziioannou, S. Tsigdinos, P. G. Tzouras, A. Nikitas and E. Bakogiannis. 'Connected and Autonomous Vehicles and Infrastructure Needs: Exploring Road Network Changes and Policy Interventions'. In: *Deception in Autonomous Transport Systems: Threats, Impacts and Mitigation Policies*. Springer, 2024, pp. 65–83.

[120] N. Kaptein and F. Claessens. 'Effects of cognitive road classification on driving behaviour: a driving simulator study'. In: (1998).

[121] A. Dijkstra. 'En route to safer roads. How road structure and road classification can affect road safety.' In: (2011).

[122] C. Gorges, K. Öztürk and R. Liebich. 'Road classification for two-wheeled vehicles'. In: *Vehicle system dynamics* 56.8 (2018), pp. 1289–1314.

[123] Y. Hu and C. Liang. 'Study on the spatial relationship between road network and the diversity of urban public facilities: the case of the central area of Changsha City'. In: *Journal of Engineering and Applied Science* 71.1 (2024), p. 156.

[124] S. Tsigdinos, G. Salamouras, I. Chatziioannou, E. Bakogiannis and A. Nikitas. 'A worldwide review of formal national street classification plans enhanced via an analytical hierarchy process: Street classification as a tool for more sustainable cities'. In: *Cities* 154 (2024a), p. 105371.

[125] J. Tumlin. *Sustainable transportation planning: tools for creating vibrant, healthy, and resilient communities*. John Wiley & Sons, 2011.

[126] I. Tang and T. P. Breckon. 'Automatic road environment classification'. In: *IEEE Transactions on Intelligent Transportation Systems* 12.2 (2010), pp. 476–484.

[127] M. Mohammadi. 'Road classification and condition determination using hyperspectral imagery'. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 39 (2012), pp. 141–146.

[128] G. Bosurgi, O. Pellegrino and G. Sollazzo. 'Road functional classification using pattern recognition techniques'. In: *The Baltic Journal of Road and Bridge Engineering* 14.3 (2019), pp. 360–383.

[129] J. Zhang, X. Chen, Y. Xiang, W. Zhou and J. Wu. 'Robust network traffic classification'. In: *IEEE/ACM transactions on networking* 23.4 (2014), pp. 1257–1270.

[130] M. Taamneh, S. Taamneh and S. Alkheder. 'Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks'. In: *International journal of injury control and safety promotion* 24.3 (2017), pp. 388–395.

[131] K. Leśniara and P. Szymański. 'Highway2vec: Representing OpenStreetMap microregions with respect to their road network characteristics'. In: *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. 2022, pp. 18–29.

[132] F. Garrido-Valenzuela, O. Cats and S. van Cranenburgh. 'Where are the people? Counting people in millions of street-level images to explore associations between people's urban density and urban characteristics'. In: *Computers, Environment and Urban Systems* 102 (2023), p. 101971.

[133] F. Garrido-Valenzuela, M. Lange, J. C. Herrera, S. van Cranenburgh and O. Cats. 'An image embedding-based approach for classifying street networks'. In: *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25)*. Minneapolis, MN, USA: ACM, 2025. DOI: 10.1145/3748636.3762715. URL: https://doi.org/10.1145/3748636.3762715.

[134] S. Tsigdinos and T. Vlastos. 'Exploring ways to determine an alternative strategic road network in a metropolitan city: A multi-criteria analysis approach'. In: *IATSS research* 45.1 (2021), pp. 102–115.

[135] D. M. Levinson and K. J. Krizek. *Planning for place and plexus: Metropolitan land use and transport*. Routledge, 2007.

[136] P. Jones and N. Boujenko. ''Link'and'Place': A new approach to street planning and design'. In: *Road & transport research: A journal of Australian and New Zealand research and practice* 18.4 (2009), pp. 38–48.

[137] B. Liu, L. Yan and Z. Wang. 'Reclassification of urban road system: integrating three dimensions of mobility, activity and mode priority'. In: *Transportation research procedia* 25 (2017), pp. 627–638.

[138] B. Immers, B. Egeter, J. Diepens and P. Westrate. *The Good Street: A new approach for rebalancing place and mobility*. 2020.

[139] P. Taylor, S. S. Anand, N. Griffiths, F. Adamu-Fika, A. Dunoyer and T. Popham. 'Road type classification through data mining'. In: *Proceedings of the 4th international conference on automotive user interfaces and interactive vehicular applications*. 2012, pp. 233–240.

[140] B. Han, D. Sun, X. Yu, W. Song and L. Ding. 'Classification of urban street networks based on tree-like network features'. In: *Sustainability* 12.2 (2020), p. 628.

[141] S. Tsigdinos, Y. Paraskevopoulos, P. G. Tzouras and K. Kepaptsoglou. 'Development of a complete method for re-conceptualizing street classification in an urban municipality'. In: *Journal of Transport Geography* 121 (2024b), p. 104025.

[142] G. Zambon, R. Benocci and G. Brambilla. 'Statistical road classification applied to stratified spatial sampling of road traffic noise in urban areas'. In: *International Journal of Environmental Research* 10.3 (2016), pp. 411–420.

[143] B. Upcroft, C. McManus, W. Churchill, W. Maddern and P. Newman. 'Lighting invariant urban street classification'. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, pp. 1712–1718.

[144] T. H. Nguyen, E. Prifti, Y. Chevaleyre, N. Sokolovska and J.-D. Zucker. 'Disease classification in metagenomics with 2d embeddings and deep learning'. In: *arXiv preprint arXiv:1806.09046* (2018).

[145] F. K. Jush, T. Truong, S. Vogler and M. Lenga. 'Medical image retrieval using pretrained embeddings'. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–5.

[146] M. Wu, Q. Huang, S. Gao and Z. Zhang. 'Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning'. In: *International Journal of Applied Earth Observation and Geoinformation* 125 (2023), p. 103591.

[147] Q. Shen, W. Zeng, Y. Ye, S. M. Arisona, S. Schubiger, R. Burkhard and H. Qu. 'StreetVizor: Visual exploration of human-scale urban forms based on street views'. In: *IEEE transactions on visualization and computer graphics* 24.1 (2017), pp. 1004–1013.

[148] Mapillary. *Mapillary*. Accessed: 2024-05-12. 2023. URL: %7Bhttps://www.mapillary.com/%7D.

[149] OSM. *Planet dump retrieved from https://osm.org*. 2024. URL: %7Bhttps://www.openstreetmap.org%7D.

[150] V. Vryniotis. *How to Train State-Of-The-Art Models Using TorchVision's Latest Primitives*. Accessed: 2025-05-04. Nov. 2021. URL: https://pytorch.org/blog/how-to-train-state-of-the-art-models-using-torchvision-latest-primitives/.

[151] A. Crespo Márquez. 'The curse of dimensionality'. In: *Digital Maintenance Management: Guiding Digital Transformation in Maintenance*. Springer, 2022, pp. 67–86.

[152] J. d. D. Ortuzar, F. J. Martinez and F. J. Varela. 'Stated preferences in modelling accessibility'. In: *International Planning Studies* 5.1 (2000), pp. 65–85.

[153] J. P. Salm, M. Bočkarjova, W. Botzen and H. Runhaar. 'Citizens' preferences and valuation of urban nature: Insights from two choice experiments'. In: *Ecological economics* 208 (2023), p. 107797.

[154] C. Green and K. Gerard. 'Exploring the social value of health-care interventions: a stated preference discrete choice experiment'. In: *Health economics* 18.8 (2009), pp. 951–976.

[155] D. Palma, J. d. D. Ortúzar, L. Rizzi and G. Casaubon. 'Modelling consumers' heterogeneous preferences: a case study with Chilean wine consumers'. In: *Australian Journal of Grape and Wine Research* 24.1 (2018), pp. 51–61.

[156] J. J. Louviere, D. A. Hensher and J. D. Swait. *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge University Press, 2000.

[157] P. Mariel, D. Hoyos, J. Meyerhoff, M. Czajkowski, T. Dekker, K. Glenk, J. B. Jacobsen, U. Liebe, S. B. Olsen, J. Sagebiel et al. *Environmental valuation with discrete choice experiments: Guidance on design, implementation and data analysis*. Springer Nature, 2021.

[158] J. Hainmueller, D. J. Hopkins and T. Yamamoto. 'Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments'. In: *Political analysis* 22.1 (2014), pp. 1–30.

[159] B. Al-Omari, J. Farhat and M. Ershaid. 'Conjoint analysis: A research method to study patients' preferences and personalize care'. In: *Journal of Personalized Medicine* 12.2 (2022), p. 274.

[160] A. G. Forster and M. Neugebauer. 'Factorial survey experiments to predict real-world behavior: A cautionary tale from hiring studies'. In: *Sociological Science* 11 (2024), pp. 886–906.

[161] P. M. Steiner, C. Atzmüller and D. Su. 'Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap'. In: *Journal of Methods and Measurement in the Social Sciences* 7.2 (2016), pp. 52–94.

[162] M. Hanemann, J. Loomis and B. Kanninen. 'Statistical efficiency of double-bounded dichotomous choice contingent valuation'. In: *American journal of agricultural economics* 73.4 (1991), pp. 1255–1263.

[163] J. M. Rose and M. C. Bliemer. 'Stated choice experimental design theory: the who, the what and the why'. In: *Handbook of choice modelling*. Edward Elgar Publishing, 2014, pp. 152–177.

[164] SurveyMonkey. *SurveyMonkey Online Surveys*. `https://www.surveymonkey.com`. Accessed: 2025-05-08. 2025.

[165] LimeSurvey. *LimeSurvey: An Open Source Survey Tool*. `https://www.limesurvey.org`. Accessed: 2025-05-08. 2025.

[166] Qualtrics. *Qualtrics Survey Platform*. `https://www.qualtrics.com`. Accessed: 2025-05-08. 2025.

[167] SurveyEngine. *SurveyEngine*. `https://surveyengine.com`. Accessed: 2025-05-08. 2025.

[168] D. L. Chen, M. Schonger and C. Wickens. 'oTree—An open-source platform for laboratory, online, and field experiments'. In: *Journal of Behavioral and Experimental Finance* 9 (2016), pp. 88–97.

[169] J. R. De Leeuw. 'jsPsych: A JavaScript library for creating behavioral experiments in a Web browser'. In: *Behavior research methods* 47 (2015), pp. 1–12.

[170] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman and J. K. Lindeløv. 'PsychoPy2: Experiments in behavior made easy'. In: *Behavior research methods* 51 (2019), pp. 195–203.

[171] F. Henninger, Y. Shevchenko, U. K. Mertens, P. J. Kieslich and B. E. Hilbig. 'lab. js: A free, open, online study builder'. In: *Behavior Research Methods* 54.2 (2022), pp. 556–573.

[172] S. Pinker. 'A theory of graph comprehension.' In: (1990).

[173] A. Vecchiato and K. Munger. 'Introducing the Visual Conjoint, with an Application to Candidate Evaluation on Social Media'. In: *Journal of Experimental Political Science* 12.1 (2025), pp. 57–71.

[174] A. López Ortega and M. Radojevic. 'Visual conjoint vs. text conjoint and the differential discriminatory effect of (visible) social categories'. In: *Political Behavior* 47.1 (2025), pp. 335–353.

[175] M. N. Hebart, C. Y. Zheng, F. Pereira and C. I. Baker. 'Revealing the multidimensional mental representations of natural objects underlying human similarity judgements'. In: *Nature human behaviour* 4.11 (2020), pp. 1173–1185.

[176] F. Garrido-Valenzuela, O. Cats and S. van Cranenburgh. 'From pixels to perceptions: using human similarity judgments to enrich urban space embeddings'. In: *Unpublished* (2025).

[177] L. I. Rizzi, J. P. Limonado and S. S. Steimetz. 'The impact of traffic images on travel time valuation in stated-preference choice experiments'. In: *Transportmetrica* 8.6 (2012), pp. 427–442.

[178] S. van Cranenburgh and F. Garrido-Valenzuela. 'Computer vision-enriched discrete choice models, with an application to residential location choice'. In: *Transportation Research Part A: Policy and Practice* 192 (2025), p. 104300.

[179] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie. 'A convnet for the 2020s'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986.

[180] PIHappiness. *What Are Image Surveys?* `https://www.pihappiness.com/what-are-image-surveys/`. Accessed: 2025-05-08. 2025.

[181] SurveyLegend. *Beautiful Image & Picture-Based Surveys*. `https://www.surveylegend.com/user-guide/beautiful-image-picture-based-surveys/`. Accessed: 2025-05-08. 2025.

[182] N. Schwitter. 'Using Artificial Intelligence to Generate Visual Vignettes in Factorial Survey Experiments'. In: (2025).

[183] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen et al. 'Promoting an open research culture'. In: *Science* 348.6242 (2015), pp. 1422–1425.

[184] M. R. Munafò, B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware and J. P. Ioannidis. 'A manifesto for reproducible science'. In: *Nature human behaviour* 1.1 (2017), p. 0021.

[185] L. K. John, G. Loewenstein and D. Prelec. 'Measuring the prevalence of questionable research practices with incentives for truth telling'. In: *Psychological science* 23.5 (2012), pp. 524–532.

[186] D. J. Leiner. *SoSci Survey*. `https://www.soscisurvey.de/`. Accessed: 2025-10-02. 2025.

[187] Questback GmbH. *Unipark / EFS Survey*. `https://www.unipark.com/`. Accessed: 2025-10-02. 2025.

[188] K. E. Train. *Discrete Choice Methods with Simulation*. 2nd. Cambridge: Cambridge University Press, 2009.

[189] R. N. Shepard. 'Multidimensional scaling, tree-fitting, and clustering'. In: *Science* 210.4468 (1980), pp. 390–398.

[190] R. Likert. 'A technique for the measurement of attitudes.' In: *Archives of psychology* (1932).

[191] R. F. DeVellis and C. T. Thorpe. *Scale development: Theory and applications*. Sage publications, 2021.

[192] Dash. *Dash: A Python Framework for Building Analytical Web Applications*. Accessed: YYYY-MM-DD. 2024. URL: `https://dash.plotly.com/`.

[193] Dash Bootstrap Components. *dash-bootstrap-components: Bootstrap components for Plotly Dash*. `https://github.com/dbc-team/dash-bootstrap-components`. Accessed: 2025-10-03, Apache-2.0 license. 2025.

[194] K. Hannum, A. M. Wellstead, M. Howlett and A. Gofen. 'Leveraging GIS for policy design: spatial analytics as a strategic tool'. In: *Policy Design and Practice* 8.1 (2025), pp. 35–49.

[195] T. Louail, M. Lenormand, O. G. Cantu Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco and M. Barthelemy. 'From mobile phone data to the spatial structure of cities'. In: *Scientific reports* 4.1 (2014), p. 5276.

[196] D. Ehrhardt, M. Behnisch, M. Jehling and M. Michaeli. 'Mapping soft densification: a geospatial approach for identifying residential infill potentials'. In: *Buildings & Cities* 4.1 (2023).

[197] S. Wan, Y. Chen, Y. Xiao, Q. Zhao, M. Li and S. Wu. 'Spatial analysis and evaluation of medical resource allocation in China based on geographic big data'. In: *BMC health services research* 21 (2021), pp. 1–18.

[198] N. Iyer, R. Menezes and H. Barbosa. 'The role of transport systems in housing insecurity: a mobility-based analysis'. In: *EPJ Data Science* 13.1 (2024), p. 49.

[199] Y. Long and J.-C. Thill. 'Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing'. In: *Computers, Environment and Urban Systems* 53 (2015), pp. 19–35.

[200] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al. 'Learning transferable visual models from natural language supervision'. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[201] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever. 'Zero-shot text-to-image generation'. In: *International conference on machine learning*. Pmlr. 2021, pp. 8821–8831.

[202] G. Collell, T. Zhang and M.-F. Moens. 'Imagined visual representations as multimodal embeddings'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[203] P. Gramacki, K. Leśniara, K. Raczycki, S. Woźniak, M. Przymus and P. Szymański. 'Srai: Towards standardization of geospatial ai'. In: *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*. 2023, pp. 43–52.

[204] W. Huang, D. Zhang, G. Mai, X. Guo and L. Cui. 'Learning urban region representations with POIs and hierarchical graph infomax'. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 196 (2023), pp. 134–145.

[205] P. Jenkins, A. Farag, S. Wang and Z. Li. 'Unsupervised representation learning of spatial data via multimodal embedding'. In: *Proceedings of the 28th ACM international conference on information and knowledge management*. 2019, pp. 1993–2002.

[206] Y. Zhang, Y. Li and F. Zhang. 'Multi-level urban street representation with street-view imagery and hybrid semantic graph'. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 218 (2024), pp. 19–32.

[207] P. Zeile, B. Resch, L. Dörrzapf, J.-P. Exner, G. Sagl, A. Summa and M. Sudmanns. 'Urban Emotions–tools of integrating people's perception into urban planning'. In: *REAL CORP 2015. PLAN TOGETHER–RIGHT NOW–OVERALL. From vision to reality for vibrant cities and regions. Proceedings of 20th international conference on urban planning, regional development and information society*. CORP–Competence Center of Urban and Regional Planning. 2015, pp. 905–912.

[208] P. Salesses, K. Schechtner and C. A. Hidalgo. 'The collaborative image of the city: mapping the inequality of urban perception'. In: *PloS one* 8.7 (2013), e68400.

[209] F. Zhang, A. Salazar-Miranda, F. Duarte, L. Vale, G. Hack, M. Chen, Y. Liu, M. Batty and C. Ratti. 'Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery'. In: *Annals of the American Association of Geographers* 114.5 (2024), pp. 876–897.

[210] T. Lorenc, S. Clayton, D. Neary, M. Whitehead, M. Petticrew, H. Thomson, S. Cummins, A. Sowden and A. Renton. 'Crime, fear of crime, environment, and mental health and wellbeing: mapping review of theories and causal pathways'. In: *Health & place* 18.4 (2012), pp. 757–765.

[211] B. E. Saelens and S. L. Handy. 'Built environment correlates of walking: a review'. In: *Medicine and science in sports and exercise* 40.7 Suppl (2008), S550.

[212] M. Zhang and A. K. Bandara. 'Understanding Pedestrians' Perception of Safety and Safe Mobility Practices'. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–17.

[213] F. Zhang, B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin and C. Ratti. 'Measuring human perceptions of a large-scale urban region using machine learning'. In: *Landscape and Urban Planning* 180 (2018), pp. 148–160.

[214] Amazon Mechanical Turk. *Amazon Mechanical Turk*. `https://www.mturk.com`. Accessed: 2025-07-01. 2025.

[215] B. J. Devereux, L. K. Tyler, J. Geertzen and B. Randall. 'The Centre for Speech, Language and the Brain (CSLB) concept property norms'. In: *Behavior research methods* 46 (2014), pp. 1119–1127.

[216] K. McRae, G. S. Cree, M. S. Seidenberg and C. McNorgan. 'Semantic feature production norms for a large set of living and nonliving things'. In: *Behavior research methods* 37.4 (2005), pp. 547–559.

[217] W. R. Tobler. 'A computer movie simulating urban growth in the Detroit region'. In: *Economic geography* 46.sup1 (1970), pp. 234–240.

[218] V. Spruyt. 'Loc2vec: Learning location embeddings with triplet-loss networks'. In: *Sentiance web article: https://www. sentiance. com/2018/05/03/venue-mapping* (2018).

[219] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell and S. Ermon. 'Tile2vec: Unsupervised representation learning for spatially distributed data'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 3967–3974.

[220] D. D. Mohan, B. Jawade, S. Setlur and V. Govindaraju. 'Deep metric learning for computer vision: A brief overview'. In: *Handbook of Statistics* 48 (2023), pp. 59–79.

[221] Y. Li, W. Huang, G. Cong, H. Wang and Z. Wang. 'Urban region representation learning with openstreetmap building footprints'. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 1363–1373.

[222] Y. Xi, T. Li, H. Wang, Y. Li, S. Tarkoma and P. Hui. 'Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests'. In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 3308–3316.

[223] A. Tversky. 'Features of similarity.' In: *Psychological review* 84.4 (1977), p. 327.

[224] E. E. Smith and D. L. Medin. *Categories and Concepts*. Cambridge, MA and London, England: Harvard University Press, 1981. DOI: doi:10.4159/harvard.9780674866270. URL: https://doi.org/10.4159/harvard.9780674866270.

[225] D. Gentner. 'Structure-mapping: A theoretical framework for analogy'. In: *Cognitive science* 7.2 (1983), pp. 155–170.

[226] P. Gardenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004.

[227] Mapillary. *Mapillary*. https://www.mapillary.com/. Accessed: 2023-11-15. 2023.

[228] Apple Inc. *Apple Maps Look Around*. Accessed: 2023-08-20. 2023. URL: https://www.apple.com/maps/.

[229] Google. *Google Street View*. Accessed: 2022-10-04. 2023. URL: https://www.google.com/maps/streetview/.

[230] Uber Technologies, Inc. *H3: A Hexagonal Hierarchical Spatial Index*. Accessed: 2023-08-20. Accessed: 2024. URL: https://h3geo.org/.

[231] Google. *Google Street View Static API*. Accessed: 2022-08-01. 2022. URL: https://developers.google.com/maps/documentation/streetview/request-streetview.

[232] CBS. *Kerncijfers per postcode*. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode. Accessed: 2023-11-02. 2023.

[233] F. I. Craik and R. S. Lockhart. 'Levels of processing: A framework for memory research'. In: *Journal of verbal learning and verbal behavior* 11.6 (1972), pp. 671–684.

[234] H. Su, S. Maji, E. Kalogerakis and E. Learned-Miller. 'Multi-view convolutional neural networks for 3d shape recognition'. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 945–953.

[235] T. Chen, S. Kornblith, K. Swersky, M. Norouzi and G. E. Hinton. 'Big self-supervised models are strong semi-supervised learners'. In: *Advances in neural information processing systems* 33 (2020), pp. 22243–22255.

[236] F. Miranda, M. Hosseini, M. Lage, H. Doraiswamy, G. Dove and C. T. Silva. 'Urban mosaic: Visual exploration of streetscapes using large-scale image data'. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–15.

[237] Q. Liang, M. Wang and T. Nagakura. 'Urban immersion: A web-based crowdsourcing platform for collecting urban space perception data'. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–8.

[238] J. Zhang, T. Piumsomboon, Z. Dong, X. Bai, S. Hoermann and R. Lindeman. 'Exploring spatial scale perception in immersive virtual reality for risk assessment in interior design'. In: *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–8.

[239] T. Vainio, I. Karppi, A. Jokinen and H. Leino. 'Towards Novel Urban Planning Methods–Using Eye-tracking Systems to Understand Human Attention in Urban Environments'. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–8.

[240] K. Sakurada, D. Tetsuka and T. Okatani. 'Temporal city modeling using street level imagery'. In: *Computer Vision and Image Understanding* 157 (2017), pp. 55–71.

[241] D. McFadden. *The measurement of urban travel demand*. Tech. rep. Institute of Urban & Regional Development, University of California, 1974.

[242] D. L. McFadden. 'Economic Choices'. In: *The American Economic Review* 91.3 (2001), pp. 351–369.

[243] G. de Jong, A. Daly, M. Pieters, C. Vellay, M. Bradley and F. Hofman. 'A model for time of day and mode choice using error components logit'. In: *Transportation Research Part E: Logistics and Transportation Review* 39.3 (2003), pp. 245–268.

[244] C. Guevara and M. Ben-Akiva. 'Endogeneity in Residential Location Choice Models'. In: *Transportation Research Record: Journal of the Transportation Research Board* 1977 (2006), pp. 60–66.

[245] S. Hess, T. Adler and J. W. Polak. 'Modelling airport and airline choice behaviour with the use of stated preference survey data'. In: *Transportation Research Part E: Logistics and Transportation Review* 43.3 (2007), pp. 221–233.

[246] C. G. Prato. 'Route choice modeling: past, present and future research directions'. In: *Journal of Choice Modelling* 2.1 (2009), pp. 65–100.

[247] A. Pinjari, R. M. Pendyala, C. R. Bhat and P. Waddell. 'Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions'. In: *Transportation* 38.6 (2011), pp. 933–958.

[248] M. J. Beck, J. M. Rose and D. A. Hensher. 'Environmental attitudes and emissions charging: An example of policy implications for vehicle choice'. In: *Transportation Research Part A: Policy and Practice* 50 (2013), pp. 171–182.

[249] S. Hess and A. Daly. *Handbook of choice modelling*. Edward Elgar Publishing, 2014.

[250] T. L. Childers, M. J. Houston and S. E. Heckler. 'Measurement of Individual Differences in Visual Versus Verbal Information Processing'. In: *Journal of Consumer Research* 12.2 (1985), pp. 125–134.

[251] T. Economist. 'How the pandemic has changed American homebuyers' preferences'. In: *The Economist* (2022).

[252] J. Lee and Y. Huang. 'Covid-19 impact on US housing markets: evidence from spatial regression models'. In: *Spatial Economic Analysis* 17.3 (2022), pp. 395–415.

[253] E. Cherchi and D. A. Hensher. 'Workshop synthesis: Stated preference surveys and experimental design, an audit of the journey so far and future research perspectives'. In: *Transportation Research Procedia* 11 (2015), pp. 154–164.

[254] P. Hevia-Koch and J. Ladenburg. 'Where should wind energy be located? A review of preferences and visualisation approaches for wind turbine locations'. In: *Energy Research & Social Science* 53 (2019), pp. 23–33.

[255] D. McFadden. 'Disaggregate Behavioral Travel Demand's RUM Side, A 30 year retrospective'. In: *Travel Behavior Research: The Leading*. 2000, pp. 17–64.

[256] S. Hess, A. Daly and R. Batley. 'Revisiting consistency with random utility maximisation: theory and implications for practical work'. In: *Theory and Decision* 84.2 (2018), pp. 181–204.

[257] P. Iglesias, M. Greene and J. d. D. Ortúzar. 'On the perception of safety in low income neighbourhoods: using digital images in a stated choice experiment'. In: *Choice Modelling: The State of the Art and the State of Practice*. 2013, pp. 193–210.

[258] R. Hurtubia, A. Guevara and P. Donoso. 'Using Images to Measure Qualitative Attributes of Public Spaces through SP Surveys'. In: *Transportation Research Procedia*. Vol. 11. 2015, pp. 460–468.

[259] B. Sifringer, V. Lurkin and A. Alahi. 'Enhancing discrete choice models with representation learning'. In: *Transportation Research Part B: Methodological* 140 (2020), pp. 236–261.

[260] I. Arkoudi, C. L. Azevedo and F. C. Pereira. 'Combining Discrete Choice Models and Neural Networks through Embeddings: Formulation, Interpretability and Performance'. In: *arXiv preprint arXiv:2109.12042* (2021). URL: https://arxiv.org/abs/2109.12042.

[261] S. van Cranenburgh, S. Wang, A. Vij, F. Pereira and J. Walker. 'Choice modelling in the age of machine learning-discussion paper'. In: *Journal of Choice Modelling* (2021), p. 100340.

[262] T. Szép, S. van Cranenburgh and C. Chorus. 'Moral rhetoric in discrete choice models: a Natural Language Processing approach'. In: *Quality & Quantity* (2023).

[263] J. Wei, W. Yue, M. Li and J. Gao. 'Mapping human perception of urban landscape from street-view images: A deep-learning approach'. In: *International Journal of Applied Earth Observation and Geoinformation* 112 (2022), p. 102886.

[264] A. Zhang, L. Song and F. Zhang. 'Perception of pleasure in the urban running environment with street view images and running routes'. In: *Journal of Geographical Sciences* 32.12 (2022), pp. 2624–2640.

[265] K. Ito, Y. Kang, Y. Zhang, F. Zhang and F. Biljecki. 'Understanding urban perception with visual data: A systematic review'. In: *Cities* 152 (2024), p. 105169.

[266] P. A. Samuelson. *Economics, an Introductory Analysis*. New York: McGraw-Hill Book Company, Inc., 1948, pp. xx + 622.

[267] R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley, 1959, p. 153.

[268] K. J. Lancaster. 'A New Approach to Consumer Theory'. In: *Journal of Political Economy* 74.2 (1966), pp. 132–157. DOI: 10.1086/259131.

[269] N. Wade and M. Swanston. *Visual perception: An introduction*. Psychology Press, 2013.

[270] M. Arriaza, J. F. Cañas-Ortega, J. A. Cañas-Madueño and P. Ruiz-Avilés. 'Assessing the visual quality of rural landscapes'. In: *Landscape and Urban Planning* 69.1 (2004), pp. 115–125. DOI: 10.1016/j.landurbplan.2003.10.029.

[271] Y. Zhao, P. E. van den Berg, I. V. Ossokina and T. A. Arentze. 'Comparing self-navigation and video mode in a choice experiment to measure public space preferences'. In: *Computers, Environment and Urban Systems* 95 (2022), p. 101828.

[272] Z. Patterson, J. M. Darbani, A. Rezaei, J. Zacharias and A. Yazdizadeh. 'Comparing text-only and virtual reality discrete choice experiments of neighbourhood choice'. In: *Landscape and Urban Planning* 157 (2017), pp. 63–74.

[273] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai and T. Chen. 'Recent advances in convolutional neural networks'. In: *Pattern Recognition* 77 (2018), pp. 354–377.

[274] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn and P. Villalobos. 'Compute trends across three eras of machine learning'. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. 2022.

[275] X. Zhai, A. Kolesnikov, N. Houlsby and L. Beyer. 'Scaling vision transformers'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[276] K. E. Train. *Discrete choice methods with simulation*. Cambridge University Press, 2003.

[277] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jégou. 'Training data-efficient image transformers & distillation through attention'. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2021.

[278] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla and M. Bernstein. 'ImageNet large scale visual recognition challenge'. In: *International Journal of Computer Vision* 115 (2015), pp. 211–252.

[279] Y. Bengio. 'Deep learning of representations for unsupervised and transfer learning'. In: *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings. 2012, pp. 17–36.

[280] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. 'Imagenet: A large-scale hierarchical image database'. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[281] B. Smith and D. Olaru. 'Lifecycle stages and residential location choice in the presence of latent preference heterogeneity'. In: *Environment and Planning A* 45.10 (2013), pp. 2495–2514.

[282] K. Ito and F. Biljecki. 'Assessing bikeability with street view imagery and computer vision'. In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103371.

[283] X. Ma, C. Ma, C. Wu, Y. Xi, R. Yang, N. Peng, C. Zhang and F. Ren. 'Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing'. In: *Cities* 110 (2021), p. 103086.

[284] F. Garrido-Valenzuela, S. van Cranenburgh and O. Cats. 'Enriching geospatial data with computer vision to identify urban environment determinants of social interactions'. In: *AGILE: GIScience Series* 3 (2022), p. 72.

[285] T. Hanibuchi, T. Nakaya and S. Inoue. 'Virtual audits of streetscapes by crowdworkers'. In: *Health & Place* 59 (2019), p. 102203.

[286] T. Tillema, B. Van Wee and D. Ettema. 'The influence of (toll-related) travel costs in residential location decisions of households: A stated choice approach'. In: *Transportation Research Part A: Policy and Practice* 44.10 (2010), pp. 785–796.

[287] K. A. Small. 'Valuation of travel time'. In: *Economics of Transportation* 1.1 (2012), pp. 2–14.

[288] K. Train and W. W. Wilson. 'Estimation of stated-preference experiments constructed from revealed-preference choices'. In: *Transportation Research Part B: Methodological* 42.3 (2008), pp. 191–203.

[289] S. van Cranenburgh, C. G. Chorus and B. Van Wee. 'Vacation behaviour under high travel cost conditions – A stated preference of revealed preference approach'. In: *Tourism Management* 43 (2014), pp. 105–118.

[290] C. A. Guevara and S. Hess. 'A control-function approach to correct for endogeneity in discrete choice models estimated on SP-off-RP data and contrasts with an earlier FIML approach by Train & Wilson'. In: *Transportation Research Part B: Methodological* 123 (2019), pp. 224–239.

[291] R. E. Nisbett and T. D. Wilson. 'Telling more than we can know: Verbal reports on mental processes'. In: *Psychological Review* 84.3 (1977), pp. 231–259.

[292] D. Ton, D. C. Duives, O. Cats, S. Hoogendoorn-Lanser and S. P. Hoogendoorn. 'Cycling or walking? Determinants of mode choice in the Netherlands'. In: *Transportation Research Part A: Policy and Practice* 123 (2019), pp. 7–23.

[293] M. Stuart-Fox, T. Kleinepier, D. Ligthart and B. Blijie. *Wonen langs de meetlat*. Tech. rep. Ministerie van Binnenlandse Zaken & Koninkrijksrelaties, Den Haag, 2022.

[294] B. H. Y. Lee and P. Waddell. 'Residential mobility and location choice: a nested logit model with sampling of alternatives'. In: *Transportation* 37.4 (2010), pp. 587–601.

[295] B. van Wee. 'Self-Selection: A Key to a Better Understanding of Location Choices, Travel Behaviour and Transport Externalities?' In: *Transport Reviews* 29.3 (2009), pp. 279–292.

[296] X. Cao. 'Examining the impacts of neighborhood design and residential self-selection on active travel: a methodological assessment'. In: *Urban Geography* (2014), pp. 1–20.

[297] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein and L. Antiga. 'PyTorch: An imperative style, high-performance deep learning library'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019.

[298] A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc., 2019.

[299] T. Hillel. 'New perspectives on the performance of machine learning classifiers for mode choice prediction: An experimental review'. In: *21st Swiss Transport Research Conference, Monte Verità, Ascona*. URL: http://www.strc.ch. 2021.

[300] J. S. Cramer. *Omitted variables and misspecified disturbances in the logit model*. Tech. rep. Discussion Paper, Manuscript. Tinbergen Institute, 2005.

[301] B. Sifringer and A. Alahi. 'Images in Discrete Choice Modeling: Addressing Data Isomorphism in Multi-Modality Inputs'. In: *arXiv preprint arXiv:2312.14724* (2023).

[302] H. Booi, W. R. Boterman and S. Musterd. 'Staying in the city or moving to the suburbs? Unravelling the moving behaviour of young families in the four big cities in the Netherlands'. In: *Population, Space and Place* 27.3 (2021), e2398.

[303] R. A. Bijker and T. Haartsen. 'More than counter-urbanisation: Migration to popular and less-popular rural areas in the Netherlands'. In: *Population, Space and Place* 18.5 (2012), pp. 643–657.

[304] H. Elshof, T. Haartsen, L. J. Van Wissen and C. H. Mulder. 'The influence of village attractiveness on flows of movers in a declining rural region'. In: *Journal of Rural Studies* 56 (2017), pp. 39–52.

[305] S. van Cranenburgh and F. Garrido-Valenzuela. 'A utility-based spatial analysis of residential street-level conditions; A case study of Rotterdam'. In: *arXiv preprint arXiv:2410.17880* (2024).

[306] A. Baevski, W. N. Hsu, Q. Xu, A. Babu and M. Gu Jin & Auli. 'Data2vec: A general framework for self-supervised learning in speech, vision and language'. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR. 2022, pp. 1298–1312.

[307] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila and F. Herrera. 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI'. In: *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

[308] Y. Liao, S. Kodagoda, Y. Wang, L. Shi and Y. Liu. 'Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks'. In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. 2016, pp. 2318–2325.

[309] S. Jansen, H. Boumeester, H. Coolen, R. Goetgeluk and E. Molin. 'The impact of including images in a conjoint measurement task: evidence from two small-scale studies'. In: *Journal of housing and the built environment* 24.3 (2009), pp. 271–297.

[310] E. Balcetis and D. Dunning. 'See what you want to see: motivational influences on visual perception.' In: *Journal of personality and social psychology* 91.4 (2006), p. 612.

[311] M. Foucault. *This is not a pipe*. University of California Press, 1983.

[312] L. Zhang, L. Zhang and X. Xu. 'Occlusion-free visualization of important geographic features in 3d urban environments'. In: *ISPRS International Journal of Geo-Information* 5.8 (2016), p. 138.

[313] K. Markvica, G. Richter and G. Lenz. 'Impact of urban street lighting on road users' perception of public space and mobility behavior'. In: *Building and environment* 154 (2019), pp. 32–43.

[314] Z. Sun, H. Jiao, H. Wu, Z. Peng and L. Liu. 'Block2vec: An approach for identifying urban functional regions by integrating sentence embedding model and points of interest'. In: *ISPRS International Journal of Geo-Information* 10.5 (2021), p. 339.

[315] B. M. Lake, T. D. Ullman, J. B. Tenenbaum and S. J. Gershman. 'Building machines that learn and think like people'. In: *Behavioral and brain sciences* 40 (2017), e253.

[316] F. Jackson. ‘What Mary didn’t know’. In: *The journal of philosophy* 83.5 (1986), pp. 291–295.

[317] R. P. Sanatani. ‘Whose Experience is it Anyway? Examining inter-subject variability in urban beauty and safety judgements’. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 47. 2025.

# Acknowledgements

I would like to begin by expressing my deepest gratitude to my supervisors, **Dr. Sander van Cranenburgh** and **Dr. Oded Cats**. Throughout these years, they provided constant guidance, thoughtful feedback, and the intellectual freedom necessary to explore new ideas and fields. Their support extended far beyond technical supervision. They helped me navigate the many highs and lows that accompany research and encouraged me to remain curious, rigorous, and persistent. I am deeply thankful for their patience, trust, and commitment throughout this journey.

I would also like to acknowledge the institutions that made this work possible. I am grateful to Delft University of Technology for providing an inspiring academic environment to develop my work. This doctoral position was made possible through the AI Lab Programme of the TU Delft AI Initiative, whose support created the opportunity to pursue this project. I also thank the Faculty of Technology, Policy and Management (TPM) for hosting me, as well as the Engineering Systems and Services (ESS) Department and the Transport and Logistics (TLO) Section for providing an intellectually stimulating environment and a welcoming academic community. In addition, I would like to thank the research school TRAIL for providing a valuable set of training courses on many relevant research topics, and for their support in the development of this thesis. Finally, I would like to thank Delft Blue, the high performance computer (HPC) from TU Delft, where I executed the AI training process of this research.

I am also grateful to several scholars and friends whose work and conversations have influenced this research in meaningful ways. I would like to thank my master supervisors, **Dr. Sebastián Raveau** and **Dr. Juan C. Herrera**. Their guidance during the earlier stages of my academic path played an important role in shaping my interest in research and ultimately encouraged me to pursue doctoral studies. Sebastián, on several occasions, listened to me, motivated me to continue, and gave me valuable guidance; Juan C. collaborated on one of the research projects that forms a part of this dissertation. I would also like to thank **Dr. Hans Löbel** for the many stimulating discussions that helped refine my ideas about deep learning; **Dr. José Ignacio Hernández** for supporting me at the beginning of my PhD journey in the Netherlands; **Dr. Ricardo Hurtubia** for those survey design suggestions and discussions on advanced DCM models; **Dr. Martin Hebart** for his inspiring work and for generously taking the time to discuss his research, which became an important reference point for one of the studies included in this dissertation; and **Dr. Remco Verzijlbergh**, who was part of my supervisory team at the beginning and gave me the right tips to start doing hands-on work during my literature review phase.

My sincere thanks also go to my colleagues and friends from the CityAI Lab: **Gabriel Nova**, the other Chilean member of the lab, with whom I discussed several points of my research and who supported me during coffee breaks and hallway chats; **Dr. Yiru Jiao** for her amazing friendship, for sharing her gastronomic culture, and for always being very supportive with my questions; **Lion Cassens** for the nice discussions about research, but also because we both share a passion for maker culture; and **Dr. Lucas Spierenburg**, who defended his PhD on the same date and with whom I have shared my entire PhD journey. We share the same passion for research, and our bond extended beyond work with some amazing board game sessions. Also, thanks to my university mates, professors, and friends: **Jeroen Delfos**, **Ali Cheshomi**, **Julien Magana**, **Joslyn Sun**, **Koichi Ito**, **Dr. Maarten Kroesen**, **Dr. Jan Anne Annema**, **Dr. Lóri Tavasszy**, **Dr. Jafar Rezaei**, and **Dr. Eric Molin**. A special thanks here to **Floris van Steijn**, who helped me translate the summary of this dissertation into Dutch. Finally, I want to thank **Conchita van der Stelt** from TRAIL, who is an exceptional person, for all the support she gave me in finishing the thesis process and throughout the journey. Overall, thank you all for sharing the daily experience of research, discussing ideas, frustrations, and discoveries made this journey far richer and more enjoyable. I am grateful to the many PhD colleagues and researchers with whom I shared conversations, coffee breaks, and moments of reflection during these years in the Netherlands. Those informal exchanges often provided clarity, perspective, and motivation when they were most needed.

I would also like to extend my thanks to all the Master and Bachelor students whom I truly enjoyed supervising, and who made my research richer in content through their contributions and discussions. A special thanks to **Roos Terra**, who expanded part of my research into her own interests. Also, **Bert Berkers**, my first Master student, who was able to mathematically advance the definition of urban embeddings. Thanks as well to all the high-quality BSc students I supervised: **Louise Heringa**, **Saul Pennings**, **Ludovica Bindi**, **Louis Claessen**, **Stijn Brouwers**, **Daniel Stiekema**, **Yair Roorda**, **Jop de Vries**, **Maarten Hulsman**, and **Max Lange**. All of them did excellent research using street-level imagery, and we had highly relevant discussions about the use of imagery in different urban applications. Some special thanks to Maarten Hulsman, who also did a Bachelor's final project under my supervision, and I truly enjoyed the learning process together with him. And a very special thanks to Max Lange for his excellent job and profound contributions to one of the chapters of this thesis.

This doctoral journey would likely not have begun without the people who brought this PhD to my attention. I am grateful to **Dr. Juan Carlos Muñoz**, **Ignacio Arismendi**, and **Francisco Proboste**, who shared the doctoral vacancy with me when it was announced. Their gesture reflected a level of trust and encouragement that I appreciate. I am also particularly thankful to Sebastián Raveau for his support and advice during the application process, which was instrumental in helping me take the first steps toward this PhD.

Beyond the academic environment, I would like to thank the friends who accompanied me throughout this stage of my life: **Dr. Jaime Soza** and **Carolina Contreras**. Their presence, conversations, and support helped maintain balance during the demanding years of doctoral research. Whether through philosophical discussions, shared experiences, or

simply moments of distraction and laughter, they contributed in ways that are difficult to fully capture in words. Additionally, special thanks to all my friends from Latin America who have surrounded me and made me feel closer to my country here in the Netherlands: **Pedro Pablo de la Barra**, **Carolina Díaz**, **Javier Pavez**, **Nicole Pereira**, **Alex Dee**, **Luis Burgos**, **Paulina Toledo**, **Tomás Burgos**, **Macarena Gaete**, **Diego Castro**, **Felipe Delgado**, **Josefina López**, **Felipe Bucci**, **Felipe López**, **Constanza Mass**, **Nicol Diaz**, **Erik Lopez**, **David Monge**, **Soledad Temporín**, **Rodrigo Tapia**, **Renzo Massobrio**, **Fernanda Berlitz**, and **Hande Öğün**. I would also like to give special thanks to my friends whom I met at the beginning of this journey, **Hanggai** and **Wilda**; we always shared our experiences and discussions, and supported each other. I also want to thank **Mauricio Orozco**, **Andrea Arévalo**, and little **Luca**. They have been present all the way through my PhD journey, sharing not only technical discussions but also a beautiful friendship. Thanks to my group of friends from PUC: **Ignacio Tiznado**, **Sebastian Muñoz**, and **Guillermo Soto**, as we have met regularly to discuss how we are doing in our lives and work. Also, thanks to **Gonzalo Cárcamo**, a friend with whom I have shared many stages of life, both in Chile and now in Europe. Finally, I would like to thank **Cristóbal Castillo**, **Pablo Vergara**, and **Tamara Guerrero** for those amazing philosophical discussions we had, which helped me become a better thinker; and **David**, **Will**, and **Kitty** my lovely former neighbors.

I am especially grateful to my family. First, I want to dedicate my deepest thanks to **Constanza Gutierrez**, my life partner, Conagu. You have always been by my side during these difficult (but amazing) years full of changes. We moved to this country and decided to start this new life together. Thank you for all the support, care, and love you give me every day. This thesis would certainly have been harder without your help and support. Also, thanks to my parents, **Eliana Valenzuela** and **Ramón Garrido**. Thank you for always encouraging me with your support, discipline, and perseverance. Your support has been constant throughout my academic journey, and I carry with me the values you instilled from an early age. A special thanks to my mother; thanks to her efforts, I was able to study in an amazing environment. Without her support and motherly love, it would have been impossible to write these acknowledgements here. Also, thanks to my brother, **Matias Garrido**, for always be present in my life, and for taking care of other responsibilities that this PhD prevented me from attending to. Finally, thanks to all my other family members and friends who are always asking about how I am doing and how my life is here in the Netherlands: **tía Nenita**, **tío Pancho**, **tío Rafa**, **Thiara**, **Eduardo Aranda**, **tía Italia**, **Fran Jesus**, **Naty**, **Faby**, **tía Sari**, **tía Bernarda**, **tía Blanca**, **tío Mario**, **Tata Mario**, **Mami Maria**, **Mami Mercedez**, **Farit**, **Vicky**, **Gaspar**, **tío Manuel**, **tío Ricardo**, **tía Tere**, **Juanjo**, **Seba**, **Pablo**, **Nico**, **tía Flora**, **Claudio Segovia**, **Davis Álvarez**, **Bárbara Torres**, **Alexis Ramírez**, **Sebastián Garcia-Huidobro**, and **Janus Leonhardt**,

Finally, I would like to dedicate a special thought to my grandparents **Francisco Valenzuela** and **Adriana Gálvez**, who are no longer with us. The values they transmitted to me, responsibility, dedication, integrity, and respect, continue to guide the way I approach both my professional and personal life. Their influence remains present in the path that ultimately led to this work.

Completing this doctoral journey has been possible thanks to the support, guidance, and encouragement of many people. To all of them, I extend my deepest gratitude.

*Francisco Orlando Garrido Valenzuela*
*Delft, April 2026*

# About the Author

**Francisco Garrido-Valenzuela** was born in 1991 in Santiago, Chile. He is an urban AI engineer and scientist whose work focuses on the intersection of cities, human perception, and data-driven methods. His research explores how people experience urban environments and how these perceptions can be incorporated into computational models to better understand and plan cities.

Francisco studied Transportation and Industrial Engineering at the Pontificia Universidad Católica de Chile. At the same university, he completed a Master of Science in Engineering, where he investigated how to incorporate technology into route choice prediction. After working in different professional and research environments, he moved to the Netherlands to pursue doctoral research at TU Delft, where he has developed new approaches to represent urban spaces using deep learning and large-scale visual data. His work aims to bridge human-centered perspectives with quantitative urban analytics, contributing to emerging methods for understanding how cities are perceived.

Beyond research, Francisco has a strong interest in technology, experimentation, and practical problem-solving. He enjoys building small technical projects—from electronics and home automation to 3D printing—that connect the digital and physical worlds. He is also interested in outdoor adventures, cycling culture, beer brewing, and coffee.

# List of publications

## Publications related to this PhD Study.

1. **F. Garrido-Valenzuela**, O. Cats, & S. van Cranenburgh, *From pixels to perceptions: Using human similarity judgments to enrich urban-space embeddings*, International Journal of Geographical Information Science (2025).

2. **F. Garrido-Valenzuela**, M. Lange, J. C. Herrera, S. van Cranenburgh, & O. Cats, *An image embedding-based approach for classifying street networks*, Proceedings of the 33rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), Minneapolis, MN, USA (2025).

3. S. van Cranenburgh & **F. Garrido-Valenzuela**, *Computer vision-enriched discrete choice models, with an application to residential location choice*, Transportation Research Part A: Policy and Practice **192**, 104300 (2025).

4. **F. Garrido-Valenzuela**, O. Cats, & S. van Cranenburgh, *Where are the people? Counting people in millions of street-level images to explore associations between people's urban density and urban characteristics*, Computers, Environment and Urban Systems **102**, 101971 (2023).

5. **F. Garrido-Valenzuela**, S. van Cranenburgh, & O. Cats, *Enriching geospatial data with computer vision to identify urban environment determinants of social interactions*, AGILE: GIScience Series **3**, 72 (2022).

## Upcoming publications related to this PhD Study.

1. S. van Cranenburgh & **F. Garrido-Valenzuela**, *Understanding environments through imagery — A landscape-science typology of information encoded in images*.

2. **F. Garrido-Valenzuela**, S. van Cranenburgh, & O. Cats, *An image embedding-based approach for classifying street networks*

3. **F. Garrido-Valenzuela**, S. van Cranenburgh, & O. Cats, *PixelSurvey: A modular Python web platform for designing and deploying image-based surveys on preferences and perceptions*

# TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 400 titles, see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Garrido-Valenzuela, F., *Pixels, People, Places: Computer Vision and Image Embeddings for Perception-Aware Urban Analytics*, T2026/10, April 2026, TRAIL Thesis Series, The Netherlands

Spierenburg, L., *Advances in the Analysis of Residential Segregation and Urban Riots*, T2026/9, April 2026, TRAIL Thesis Series, The Netherlands

Boot, M., *Evaluating Experiences with Smart Cycling Technologies: Sensor-based evaluations of outdoor cycling experiences with Smart Cycling Technologies*, T2026/8, March 2026, TRAIL Thesis Series, The Netherlands

Wen, X., *Data-Driven Spatial-Temporal Modeling for Bicycle Traffic Prediction*, T2026/7, March 2026, TRAIL Thesis Series, The Netherlands

Wang, Z., *Optimising Performance of Automatic Train Operation on Railway Networks*, T2026/6, March 2026, TRAIL Thesis Series, The Netherlands

Hadi, A.H., *DEM Modelling of Multi-Component Segregation in the Blast Furnace Charging System*, T2026/5, February 2026, TRAIL Thesis Series, The Netherlands

Farhani, M., *Demand Management Strategies for Operations of Shared Mobility Services*, T2026/4, February 2026, TRAIL Thesis Series, The Netherlands

Yao, X., *Driving Heterogeneity in Traffic Flow Theory: An action-based framework for identification, modelling, and simulation*, T2026/3, January 2026, TRAIL Thesis Series, The Netherlands

Versluis, N.D., *Optimising Railway Traffic Management under Radio-Based Distance-to-Go Signalling*, T2026/2, January 2026, TRAIL Thesis Series, The Netherlands

Jiao, Y., *Proactive Collision Risk Quantification in Multi-directional Traffic Interactions*, T2026/1, January 2026, TRAIL Thesis Series, The Netherlands

Asadi, M., *Accessibility and Road Safety: Integration of road safety in accessibility evaluation*, T2025/20, November 2025, TRAIL Thesis Series, The Netherlands

Akse, R., *Understanding and untangling the uncertainty knot: How to catalyse decision-making in mobility innovations*, T2025/19, November 2025, TRAIL Thesis Series, The Netherlands

Führer, K., *Participatory Decision-making under Deep Uncertainty: Modelling mobility*

*transitions*, T2025/18, November 2025, TRAIL Thesis Series, The Netherlands

Picco, A., *Monitoring and Feedback in Driving*, T2025/17, October 2025, TRAIL Thesis Series, The Netherlands

Cebeci, M.S., *Behaviour of Prosumers in Last-mile Logistics: The case of crowdshipping*, T2025/16, September 2025, TRAIL Thesis Series, The Netherlands

Kuijpers, A., *Enabling Inter-Organizational Collaboration Through Platforms: The role of trust*, T2025/15, September 2025, TRAIL Thesis Series, The Netherlands

Song, R., *Human-MASS Interaction in Decision-Making for Safety and Efficiency in Mixed Waterborne Transport System*, T2025/14, June 2025, TRAIL Thesis Series, The Netherlands

Destyanto, A.R., *A Method for Evaluating Port Resilience in an Archipelago*, T2025/13, June 2025, TRAIL Thesis Series, The Netherlands

Karademir, C., *Synchronized Two-echelon Routing Problems: Exact and approximate methods for multimodal city logistics*, T2025/12, May 2025, TRAIL Thesis Series, The Netherlands

Vial, A., *Eyes in Motion: A new traffic sensing paradigm for pedestrians and cyclists*, T2025/11, May 2025, TRAIL Thesis Series, The Netherlands

Chen, Q., *Towards Mechanical Intelligence in Soft Robotics: Model-based design of mechanically intelligent structures*, T2025/10, April 2025, TRAIL Thesis Series, The Netherlands

Eftekhar, Z., *Exploring the Spatial and Temporal Patterns in Travel Demand: A data-driven approach*, T2025/9, June 2025, TRAIL Thesis Series, The Netherlands

Reddy, N., *Human Driving Behavior when Interacting with Automated Vehicles and the Implications on Traffic Efficiency*, T2025/8, May 2025, TRAIL Thesis Series, The Netherlands

Durand, A., *Lost in Digitalisation? Navigating public transport in the digital era*, T2025/7, May 2025, TRAIL Thesis Series, The Netherlands

Dong, Y., *Safe, Efficient, and Socially Compliant Automated Driving in Mixed Traffic: Sensing, Anomaly Detection, Planning and Control*, T2025/6, May 2025, TRAIL Thesis Series, The Netherlands

Droffelaar, I.S. van, *Simulation-optimization for Fugitive Interception*, T2025/5, May 2025, TRAIL Thesis Series, The Netherlands

Fan, Q., *Fleet Management Optimisation for Ride-hailing Services: from mixed traffic to fully automated environments*, T2025/4, April 2025, TRAIL Thesis Series, The Netherlands

Hagen, L. van der, *Machine Learning for Time Slot Management in Grocery Delivery*, T2025/3, March 2025, TRAIL Thesis Series, The Netherlands