TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Garrido Valenzuela, F. O., 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 (online)*, *40*(7), 2421-2453. https://doi.org/10.1080/13658816.2025.2595658

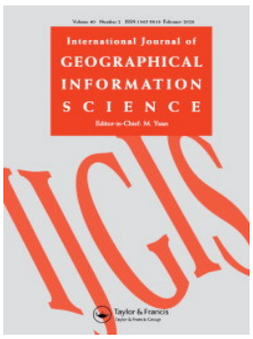
Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.



International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

From pixels to perceptions: using human similarity judgments to enrich urban space embeddings

Francisco Garrido-Valenzuela, Oded Cats & Sander van Cranenburgh

To cite this article: Francisco Garrido-Valenzuela, Oded Cats & Sander van Cranenburgh (03 Feb 2026): From pixels to perceptions: using human similarity judgments to enrich urban space embeddings, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2595658

To link to this article: https://doi.org/10.1080/13658816.2025.2595658

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 03 Feb 2026.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tgis20

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2595658

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

From pixels to perceptions: using human similarity judgments to enrich urban space embeddings

Francisco Garrido-Valenzuela[a,b], Oded Cats[b] and Sander van Cranenburgh[a]

[a]CityAI Lab, Transport and Logistics Group, Delft University of Technology, Delft, The Netherlands; [b]Department of Transport & Planning, Delft University of Technology, Delft, The Netherlands

ABSTRACT

This research introduces a new method for constructing and training an Urban Space Embedding Model (USEM) by integrating human perceptions and street-level images (SLI) into its formulation. Traditional urban embedding models often overlook subjective human experiences, such as perceptions of safety or attractiveness. To address this gap, our method leverages similarity judgments from over 1500 participants, who compared different urban spaces based on SLI. These human judgments were then used as a supervision signal in training the USEM, allowing the model to capture both visual and perceptual information about urban spaces. The method is implemented across the Netherlands, using around one million geo-tagged SLI, and demonstrated in Rotterdam. This approach represents a significant advancement in urban computing by incorporating human-centered data into urban modeling. It offers new opportunities for city planners and policymakers to better understand how urban spaces are perceived and to consider these perceptions in efforts to design more livable and inclusive environments.

ARTICLE HISTORY
Received 20 March 2025
Accepted 22 November 2025

KEYWORDS
Human similarity judgment; urban representation learning; street-level images; computer vision; human perceptions

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 (Hannum *et al.* 2025) and support evidence-based decision-making (Louail *et al.* 2014). 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 (Ehrhardt *et al.* 2023), optimize resource allocation (Wan *et al.* 2021), and address demanding urban challenges (Long and Thill 2015, Iyer *et al.* 2024). 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

CONTACT Francisco Garrido-Valenzuela f.garridov@uc.cl

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

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 (Collell *et al.* 2017, Radford *et al.* 2021, Ramesh *et al.* 2021). 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 (Wang *et al.* 2020, Huang *et al.* 2021, Woźniak and Szymański 2021, Gramacki *et al.* 2023). 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 (Huang *et al.* 2023), travel demand flows from GPS traces (Jenkins *et al.* 2019), and visual content from street-level images (SLI) (Zhang *et al.* 2024b) 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 (Zeile *et al.* 2015). 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 (Salesses *et al.* 2013, Naik *et al.* 2014, Dubey *et al.* 2016), and a comprehensive overview of visual-intelligence approaches is provided by Zhang *et al.* (2024a). These subjective aspects are relevant for policy making, as they provide people-centered insight into the design of public spaces and urban policies (Ramírez *et al.* 2021). For instance, perceived safety in neighborhoods affects people's mental well-being (Lorenc *et al.* 2012) and correlates with the amount of physical activity (Saelens and Handy 2008, Zhang and Bandara 2024). In addition, Abass and Tucker (2021) 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 (Zhang *et al.* 2018) 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 crowdsourcing (e.g., Amazon Mechanical Turk

(2025)) 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 (Dubey *et al.* 2016). 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 (Hebart *et al.* 2020). 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 (McRae *et al.* 2005, Devereux *et al.* 2014). 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 Garrido-Valenzuela *et al.* (2023). 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 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 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.

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.

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 Mikolov (2013), 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 1 conceptualizes the process and goal of creating an urban embedding.

This process is formalized as the *Urban Neighborhood Embedding Problem* (Huang *et al.* 2021), 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

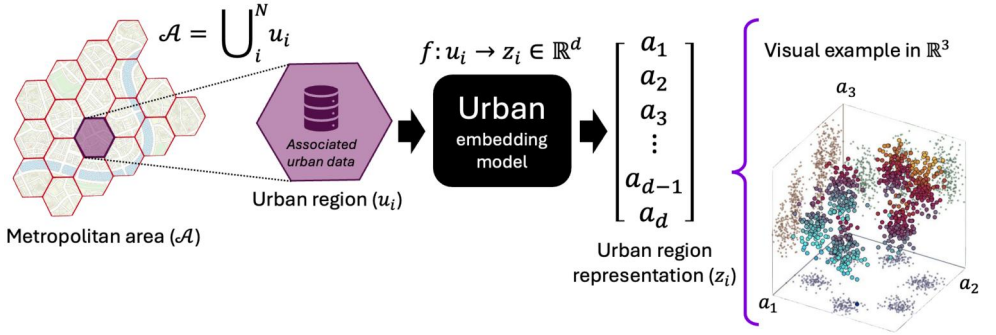


Figure 1. General procedure for creating an urban space embedding in the urban representation field. A metropolitan area $\mathcal{A}$ is subdivided into a set of urban regions u_i. All data in u_i is processed and mapped into $\mathbb{R}^d$. A visual example in $\mathbb{R}^3$ is shown. Notation is based on Huang *et al.* (2021).

patterns, assess the impact of urban interventions, predict socioeconomic indicators in data-limited areas, or support location-based services and recommendations.

2.1.1. 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 (Tobler 1970). 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 (Woźniak and Szymański 2021) uses Tobler's law for sampling urban regions and employs the Skip-gram model with negative sampling (Mikolov 2013). 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 (Spruyt 2018), Tile2Vec (Jean *et al.* 2019), and Urban2Vec (Wang *et al.* 2020) use Tobler's law for sampling regional instances. These models sampled triplets of regions for use with the triplet margin loss (Schroff *et al.* 2015). 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 (Mohan *et al.* 2023). Alternative approaches to urban embedding do not rely directly on Tobler's Law. For instance, the RegionEncoder model (Jenkins *et al.* 2019) 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.

2.1.2. 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 (Huang *et al.* 2021).

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, Li *et al.* (2023) 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, Xi *et al.* (2022) combined satellite images with POIs, adding a bird's eye visual component to the function provided by the amenities. Urban2Vec (Wang *et al.* 2020) integrated multi-modal data, including Street View images and textual data related to POIs, to enhance the contextual understanding of urban neighborhoods. Additionally, Huang *et al.* (2021) 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.

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 (Tversky 1977). For instance, a knife can be perceived as a weapon with a dangerous connotation, or as a utensil with a practical connotation (McRae *et al.* 2005). These judgments are grounded in cognitive processes that evaluate similarities and differences among various stimuli based on human perceptions and experiences (Smith and Medin 1981). 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 (Gentner 1983).

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 (Gardenfors 2004), 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 (Dubey *et al.* 2016) 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.

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. Hebart *et al.* (2020) 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, Hebart *et al.* (2020) 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.

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.

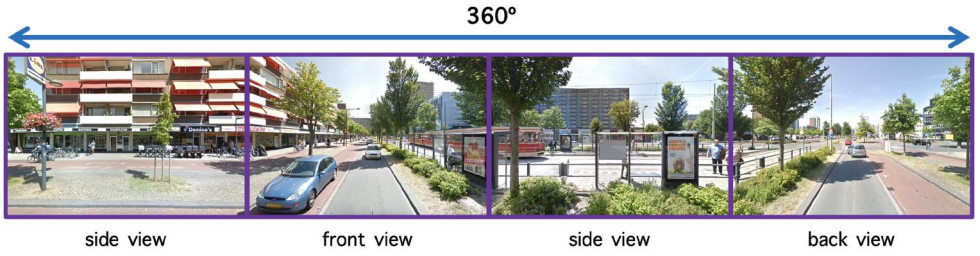


Figure 2. Decomposition of one 360-degree image into four individual 90-degree images. Figure from Garrido-Valenzuela *et al.* (2023).

3.1. Required data

3.1.1. 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 (Mapillary 2023), and Apple Look Around (Apple Inc. 2023), which make available these types of images around the globe. In this study, we use street-level images from GSV (Google 2023), following the image ID collection method proposed by Garrido-Valenzuela *et al.* (2023). 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 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.

3.1.2. 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 (Uber Technologies, Inc. 2024). 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 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.

3.1.3. 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.

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 (Google 2022), following the procedure described in Garrido-Valenzuela *et al.* (2023). 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 (CBS 2023), 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.

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* (Huang *et al.* 2021) by incorporating human perceptions into its formulation. Figure 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.

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

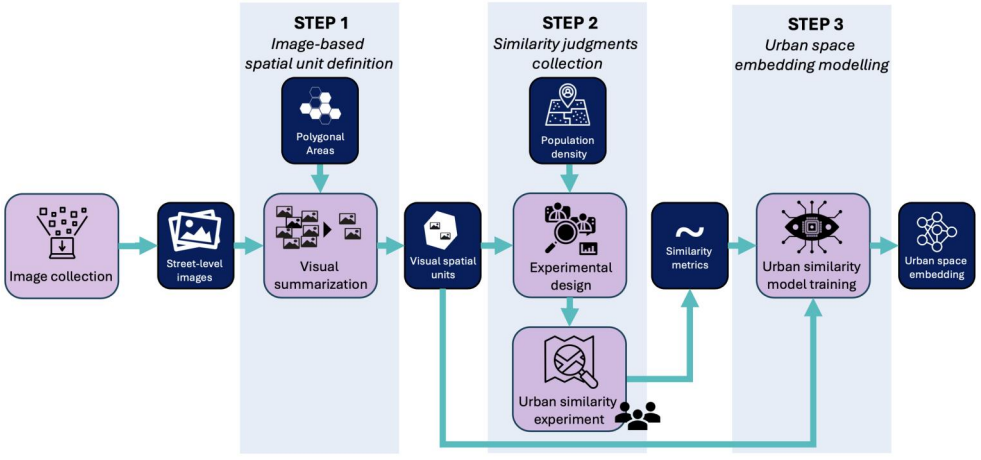


Figure 3. General pipeline of the method. Purple (large) boxes are the key modeling processes in each step, and blue (small) boxes are the input/output of each sub-step. In data collection, images are retrieved. In step 1, the sampled images are spatially associated with spatial units. Then, in step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

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 4 shows the procedure for obtaining k subset of images from a spatial unit.

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 (He *et al.* 2016), 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 5 illustrates the results of our summarization

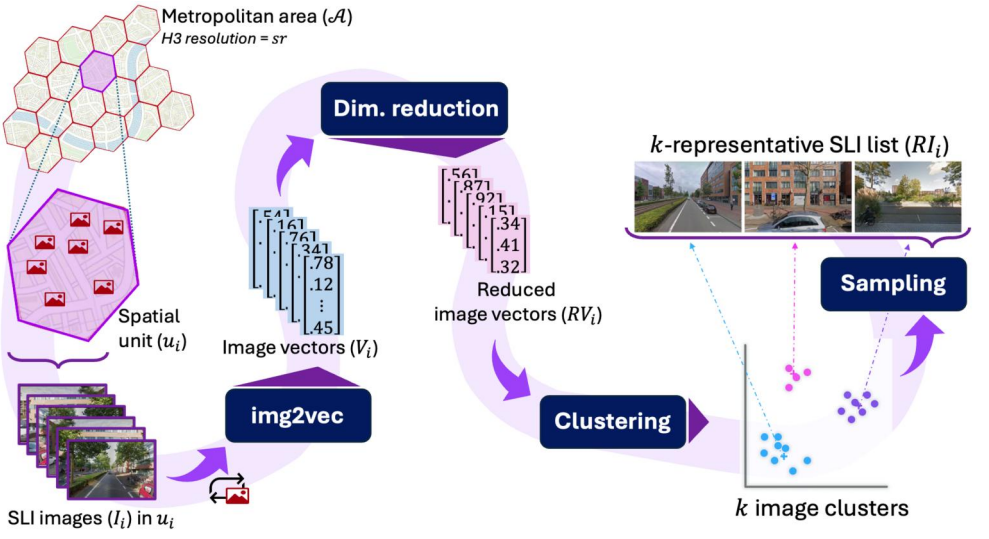


Figure 4. Visual summarization algorithm used for sampling a diverse subset of k images from each spatial unit. The images associated with a polygonal area are converted into vectors using any img2vec model. Then, the vectors are reduced using a dimensionality reduction technique and clustered into k classes. Finally, one image per cluster is sampled.

algorithm, showcasing how the original set of images within a spatial unit is distilled into five diverse and representative images.

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.

4.2.1. 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

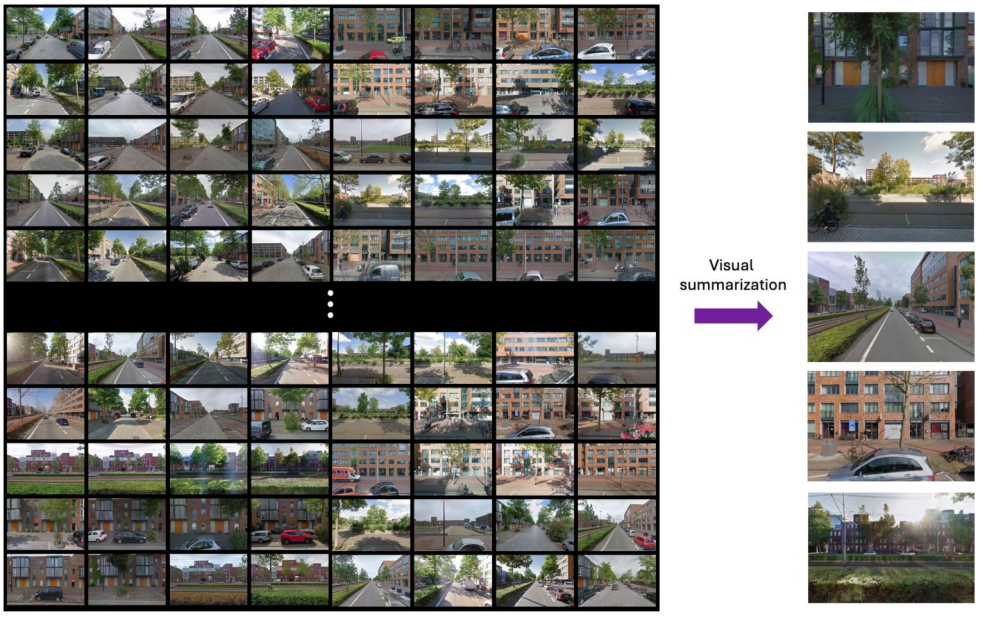


Figure 5. Application of the visual summarization algorithm in a spatial unit with more than 100 images.

Margin Loss introduced by Schroff *et al.* (2015), 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 (Craik and Lockhart 1972).

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(a). Next, we divide the range of log-transformed population density into five ranges, as shown in Figure 6(a) 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

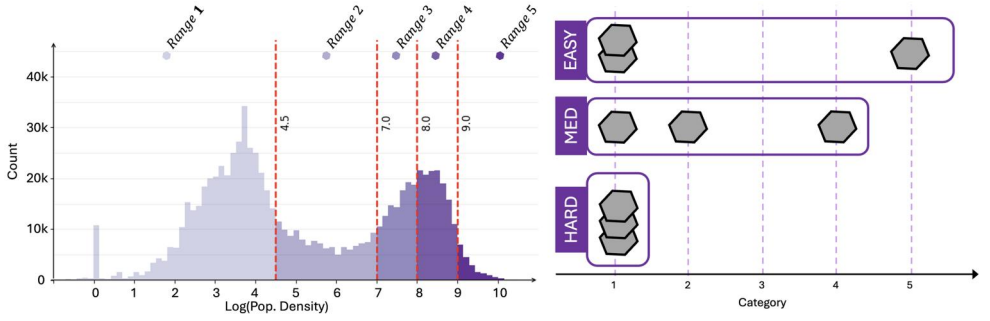


Figure 6. (a) Histogram of the Log(population density) in the Netherlands. Vertical dotted lines divide the spectrum into five ranges of population density. (b) Examples of triplet tasks with different difficulties based on the population density ranges of the spatial units.

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.

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 Eqs. 1–3. The ranges come from the splits made on the population density distribution (see Figure 6(a). 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(b). 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(b).

$$minDist = \min(|R_2 - R_1|, |R_3 - R_1|, |R_3 - R_2|) \tag{1}$$

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

$$difficulty = Rank\left(\frac{minDist + 0.1}{avgMaxDist + 0.1}\right) \tag{3}$$

Mathematically, the difficulty of a task is determined by Eq. 3, where *minDist* (Eq. 1) corresponds to the smallest difference in population density range indices among the three pairs of places in the triplet; and *avgMaxDist* (Eq. 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]).

4.2.2. 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 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 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.

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 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

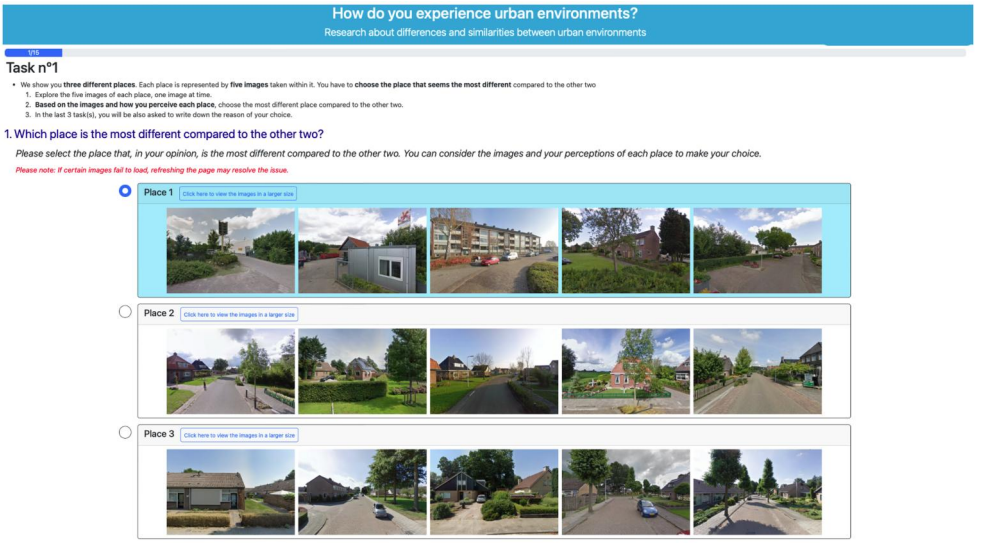


Figure 7. Interface of the web platform for the urban similarity experiment. Participants are presented with three spatial units and five images for each of which, and are asked to identify the most different spatial unit.

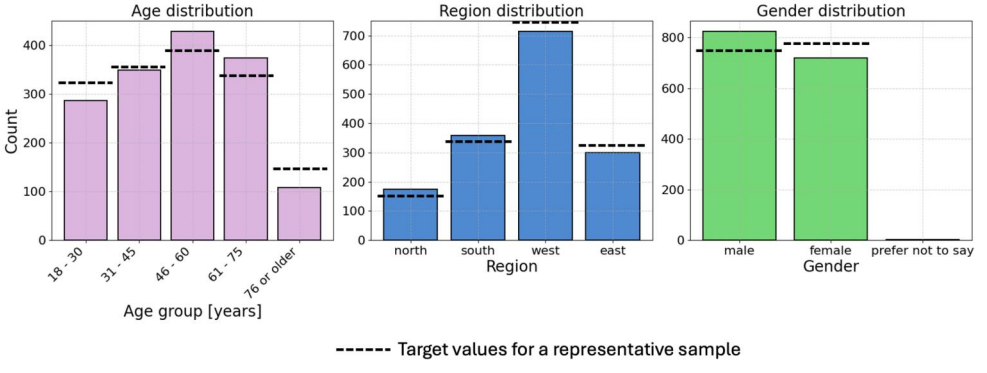


Figure 8. Demographic distribution of the participants in the urban similarity experiment.

to learn the similarity metrics provided by participants through the optimization of a triplet margin loss function (Schroff *et al.* 2015). 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.

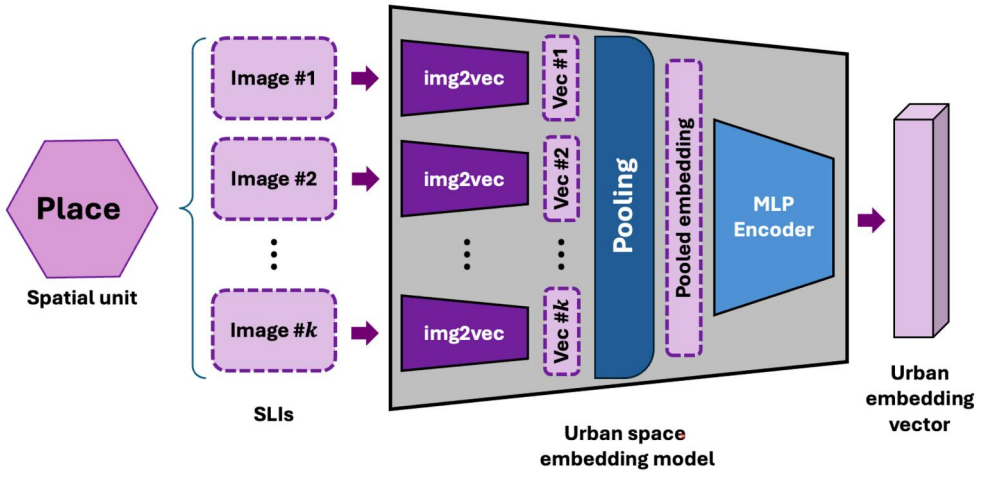


Figure 9. Urban space embedding model architecture. Each spatial unit is represented by k images (three in the figure), processed by an img2vec model, pooled, and encoded by MLP layers to produce the final embedding vector.

4.3.1. 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 Su *et al.* (2015) for processing 3D objects using a set of images, and are adopted here as each spatial unit is represented by $k = 5$ images. Figure 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 (He *et al.* 2016) and ConvNext (Liu *et al.* 2022) 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.

4.3.2. 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

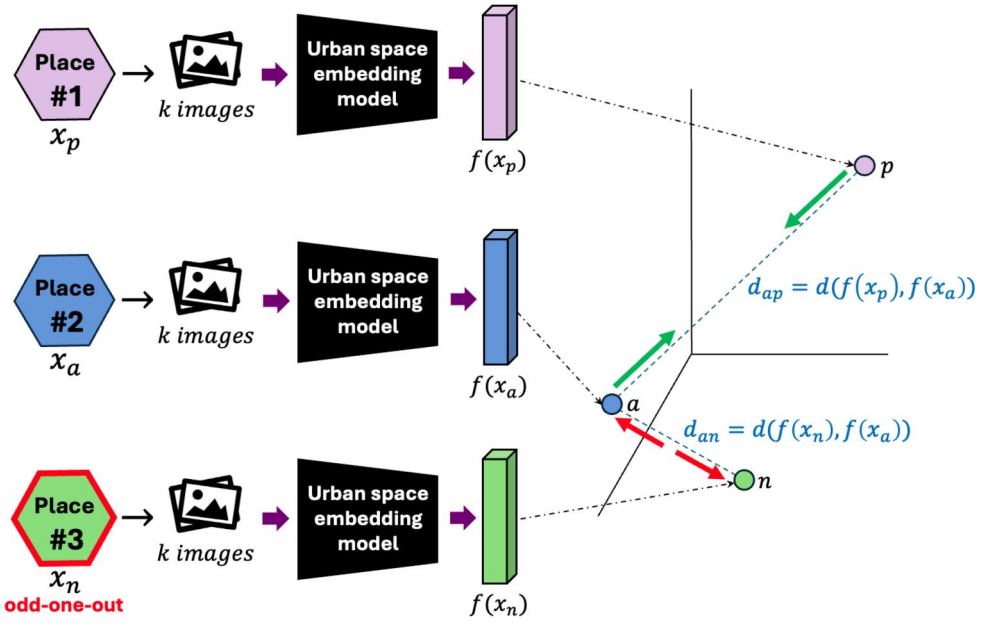


Figure 10. Triplet network architecture for training the embedding model. The training is based on the experiment responses (odd-one-outs), illustrated by place #3 in the figure. Each place is represented by *k* images. These *k* images are mapped into an embedding space.

runners, we obtain a training set with 16,654 (90%) triplets and a testing set with 1,852 (10%) triplets.

4.3.3. Triplet network architecture for training the USEM

We employ a triplet network architecture for training the USEM presented in Figure 9. A triplet network consists of three identical sub-networks, each processing one of the three spatial units within a triplet. Figure 10 schematically presents this architecture, where each sub-network (black trapezoid) is the USEM (i.e., Figure 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 (Schroff *et al.* 2015), with the odd-one-out place selected by the respondent as the negative instance.

Figure 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

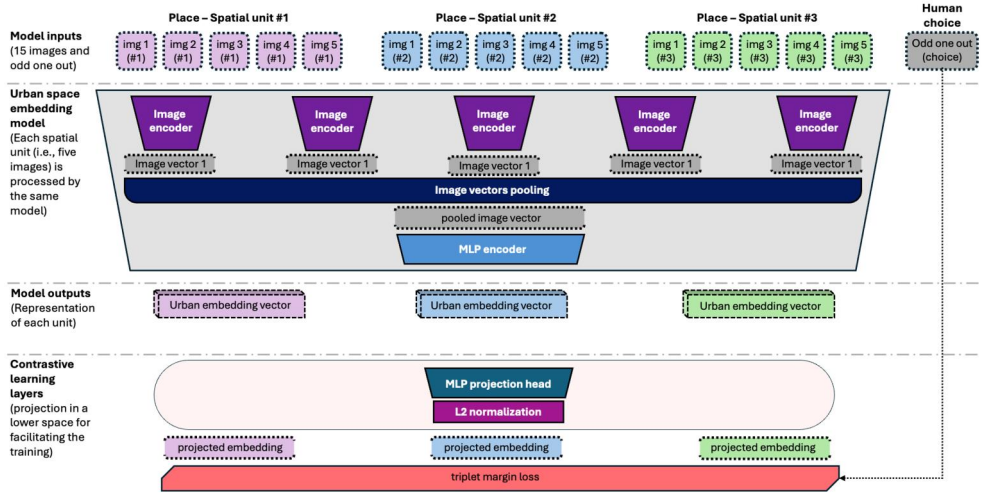


Figure 11. Architecture of the triplet network used for training the urban space embedding model. The embedding model processes the five images of each spatial unit and generates an urban embedding vector. Then, the contrastive learning layers map the embeddings in a unit hypersphere for computing the margin loss and updating the parameters of the model.

of dimensionality, and projecting the vectors into a lower-dimensional space can improve the model's performance (Chen *et al.* 2020). The L2 normalization layer was adopted from the FaceNet architecture (Schroff *et al.* 2015) 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.

4.3.4. Training procedure

For training the full model, we employ the triplet margin loss function as defined by Schroff *et al.* (2015). 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 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 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 Eq. 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).

$$||f(x_a) - f(x_p)||^2 + \alpha < ||f(x_a) - f(x_n)||^2 \tag{4}$$

In Eq. 4, $f(x)$ represents the embedding of a spatial unit x, and α 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 Eq. 5.

$$L = \sum_{i=1}^{N} \max(0, ||f(x_a^{(i)}) - f(x_p^{(i)})||^2 - ||f(x_a^{(i)}) - f(x_n^{(i)})||^2 + \alpha) \quad (5)$$

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.

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 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.

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.

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

Table 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

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.* (2020) 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.

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 generalises 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.

5.2.1. 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 (Woźniak and Szymański 2021). 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 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*),

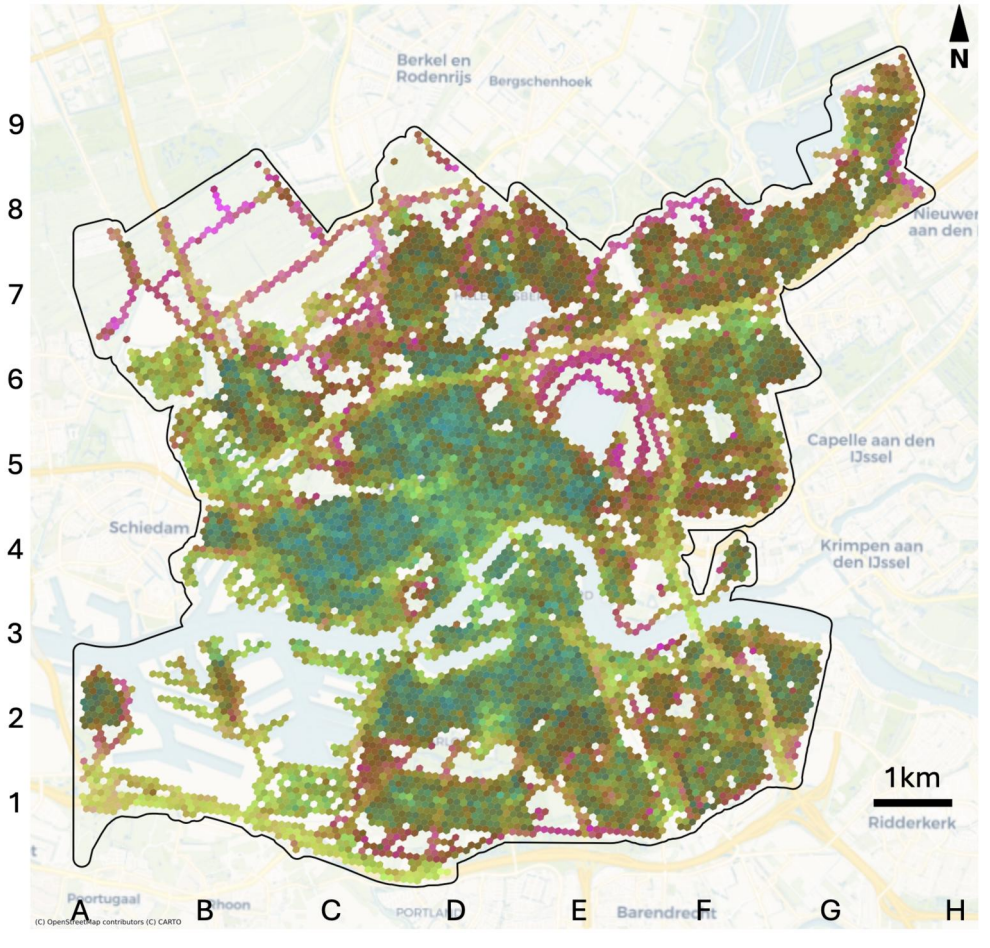


Figure 12. Embedding values projected in three dimensions using PCA. Each hexagon's color is produced by normalizing the PCA values between 0 and 255 and using them as RGB channels. Similar colors indicate similar embeddings in the PCA space, and therefore similar urban spaces.

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.

5.2.2. 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 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 12. For instance, the city center hexagon (Figure 13(a)) shows high similarity with hexagons in the immediate vicinity, such as the adjacent hexagons in the city center and some clusters located in the suburban areas (which are also light-colored in Figure 12). The periphery hexagon (Figure 13(b)) has a low similarity with the city center, but a higher similarity with other periphery hexagons. The park hexagon (Figure 13(c)) is most similar to hexagons in the northeast area, where green spaces are prevalent. Finally, the highway hexagon (Figure 13(d)) 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 13(b)) and the city center hexagon (Figure 13(a)) 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.

5.2.3. Zonification of spatial units

We also delineate different zones based on the proximity of units and embedding features. While the map in Figure 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

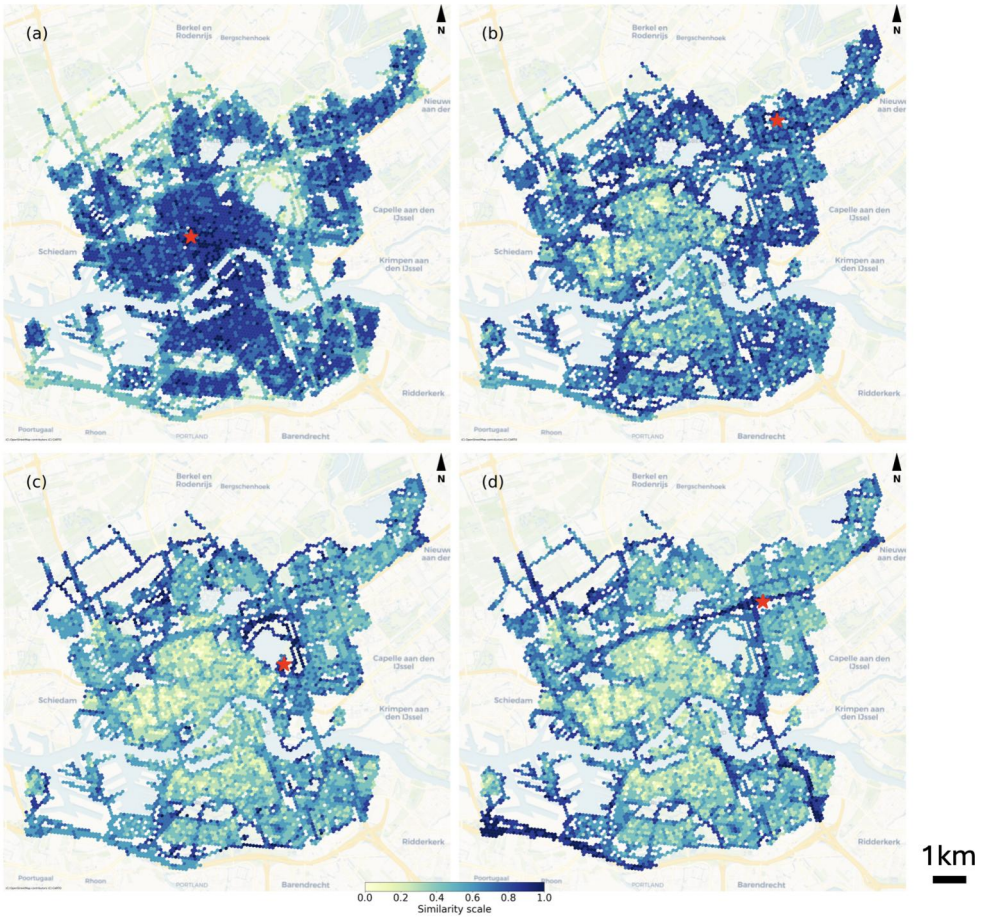


Figure 13. Heatmaps showing the similarities between Rotterdam's urban spaces using four reference hexagons. Each map represents the similarity of one hexagon (marked by the white star) with all others in Rotterdam. (a) hexagon in the city center, (b) hexagon in the periphery, (c) hexagon in a park, and (d) hexagon in a highway.

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 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 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 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 12 and 13(c), 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 arterials are clustered together in Zone E, depicted in grey, 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.

5.2.4. 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 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 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

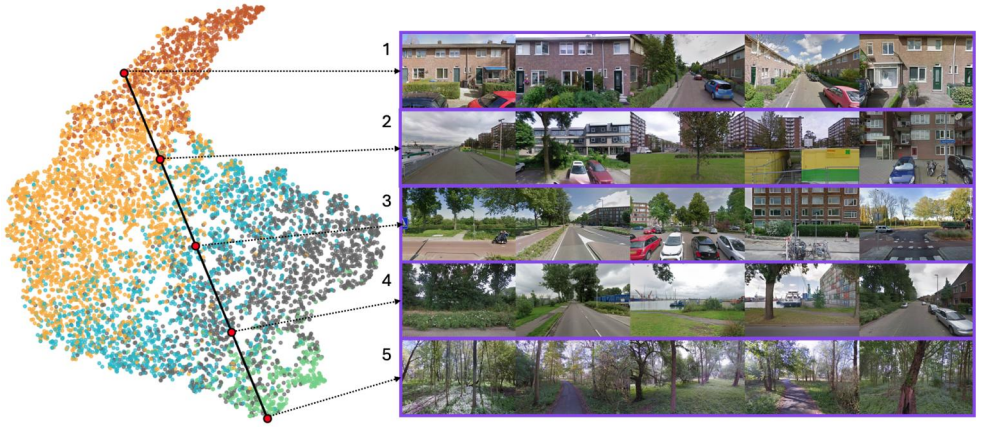


Figure 15. Exploration of the multidimensional urban space embeddings in Rotterdam. The left side shows the t-SNE projection of the embeddings colored by the identified zones. The right side displays images from hexagons along a path connecting two random hexagons in the embedding space.

points along this path and display the corresponding images. Figure 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 15 effectively groups the zones described in Figure 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, grey (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 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.

5.2.5. 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

Figure 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.

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 16 displays the clusters identified using the ImageNet model. We try to replicate the colors used in Figure 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 14, to which this model assigns a unique cluster. The image samples include zones B and C (see Figure 14).

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 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 grey). 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 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. 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 Miranda *et al.* (2020) have demonstrated the potential of large-scale visual data in exploring and differentiating streetscapes, while Zhang and Bandara (2024) 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 (Hebart *et al.* 2020).

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, ambience, 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 (Schroff *et al.* 2015), which often led to a collapsed model where the loss converged to the margin value, resulting in sparse embedding vectors (Schroff *et al.* 2015). To address this, we employed larger batch sizes (Chen *et al.* 2020) and incorporated L2 projection before computing the loss (Schroff *et al.* 2015), 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 Liang *et al.* (2020) and Zhang *et al.* (2020) have explored the potential of VR-based platforms for capturing perceptions of urban spaces, while Vainio *et al.* (2019) 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 Sakurada *et al.* (2017), 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.

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.

Declaration of generative AI in the writing process

During the preparation of this work the authors used chatGPT in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the TU Delft AI Labs programme.

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/TUD-CityAI-Lab/From-pixels-to-perceptions

Notes on contributors

Francisco Garrido-Valenzuela is a PhD candidate in CityAI Lab, at TU Delft, in the Netherlands. His research focuses on studying cities using spatial and AI methods. He contributed to the conceptualization, data collection, experiment implementation, modeling and training, and manuscript preparation.

Oded Cats is Professor in the Department of Transport and Planning and Director of the CityAI Lab, at TU Delft, in the Netherlands. His research interests lie in the intersection between transport networks, operations, policy and travel behavior. He is the PhD supervisor of this study.

Sander van Cranenburgh is Associate Professor and Director of the CityAI Lab, at TU Delft, in the Netherlands. His research focuses on developing new models for enhancing our understanding of human choice behavior. He is the PhD supervisor of this study.

References

Abass, Z.I., and Tucker, R., 2021. Talk on the street: The impact of good streetscape design on neighbourhood experience in low-density suburbs. *Housing, Theory and Society*, 38 (2), 204–227.

Amazon Mechanical Turk, 2025. Amazon Mechanical Turk. https://www.mturk.com [Accessed 1 July 2025].

Apple Inc., 2023. Apple maps look around. Available from: https://www.apple.com/maps/ [Accessed: 20 Aug 2023].
CBS, 2023. Kerncijfers per postcode. Available from: https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode [Accessed 2 Nov 2023].
Chen, T., *et al.*, 2020. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33, 22243–22255.
Collell, G., Zhang, T., and Moens, M.F., 2017. Imagined visual representations as multimodal embeddings. *In: Proceedings of the AAAI conference on artificial intelligence*. vol. 31.
Craik, F.I., and Lockhart, R.S., 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11 (6), 671–684.
Devereux, B.J., *et al.*, 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior Research Methods*, 46 (4), 1119–1127.
Dubey, A., *et al.*, 2016. Deep learning the city: Quantifying urban perception at a global scale. *In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 196–212.
Ehrhardt, D., *et al.*, 2023. Mapping soft densification: a geospatial approach for identifying residential infill potentials. *Buildings & Cities*, 4 (1), 193–211.
Gardenfors, P., 2004. *Conceptual spaces: The geometry of thought*. MIT Press.
Garrido-Valenzuela, F., Cats, O., and 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.
Gentner, D., 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7 (2), 155–170.
Google, 2022. Google Street View Static API. Available from: https://developers.google.com/maps/documentation/streetview/request-streetview. [Accessed 1 Aug 2022].
Google, 2023. Google street view. Available from: https://www.google.com/maps/streetview/. [Accessed 4 Oct 2022].
Gramacki, P., *et al.*, 2023. SRAI: Towards standardization of geospatial AI. *In: Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI '23)*. New York, NY: Association for Computing Machinery, 43–52.
Hannum, K., *et al.*, 2025. Leveraging gis for policy design: spatial analytics as a strategic tool. *Policy Design and Practice*, 8 (1), 35–49.
He, K., *et al.*, 2016. Deep residual learning for image recognition. *In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.
Hebart, M.N., *et al.*, 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4 (11), 1173–1185.
Huang, T., *et al.*, 2021. M3g: Learning urban neighborhood representation from multi-modal multi-graph. *Proceedings of the DeepSpatial*, 2021.
Huang, W., *et al.*, 2023. Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196, 134–145.
Iyer, N., Menezes, R., and Barbosa, H., 2024. The role of transport systems in housing insecurity: a mobility-based analysis. *EPJ Data Science*, 13 (1), 49.
Jean, N., *et al.*, 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01), 3967–3974.
Jenkins, P., *et al.*, 2019. Unsupervised representation learning of spatial data via multimodal embedding. *In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. New York, NY: Association for Computing Machinery, 1993–2002.
Li, Y., *et al.*, 2023. Urban region representation learning with openstreetmap building footprints. *In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. New York, NY: Association for Computing Machinery, 1363–1373.
Liang, Q., Wang, M., and Nagakura, T., 2020. 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 (CHI EA '20). New York, NY: Association for Computing Machinery, 1–8.
Liu, Z., *et al.*, 2022. A convnet for the 2020s. *In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 11976–11986.
Long, Y., and Thill, J.C., 2015. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*, 53, 19–35.
Lorenc, T., *et al.*, 2012. Crime, fear of crime, environment, and mental health and wellbeing: mapping review of theories and causal pathways. *Health & Place*, 18 (4), 757–765.
Louail, T., *et al.*, 2014. From mobile phone data to the spatial structure of cities. *Scientific Reports*, 4 (1), 5276.
Mapillary, 2023. Mapillary. https://www.mapillary.com/ [Accessed 15 Nov 2023].
McRae, K., *et al.*, 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37 (4), 547–559.
Mikolov, T., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://arxiv.org/abs/1301.3781
Miranda, F., *et al.*, 2020. Urban mosaic: Visual exploration of streetscapes using large-scale image data. *In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. New York, NY: Association for Computing Machinery, 1–15.
Mohan, D.D., *et al.*, 2023. Deep metric learning for computer vision: a brief overview. *Handbook of Statistics*, 48, 59–79.
Naik, N., *et al.*, 2014. Streetscore-predicting the perceived safety of one million streetscapes. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 779–785.
Radford, A., *et al.*, 2021. Learning transferable visual models from natural language supervision. *In: Proceedings of the 38th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*, vol. 139, 8748–8763.
Ramesh, A., *et al.*, 2021. Zero-shot text-to-image generation. *In: Proceedings of the 38th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*, vol. 139, 8821–8831.
Ramírez, T., *et al.*, 2021. Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety. *Landscape and Urban Planning*, 208, 104002.
Saelens, B.E., and Handy, S.L., 2008. Built environment correlates of walking: a review. *Medicine and Science in Sports and Exercise*, 40 (7 Suppl), S550–S566.
Sakurada, K., Tetsuka, D., and Okatani, T., 2017. Temporal city modeling using street level imagery. *Computer Vision and Image Understanding*, 157, 55–71.
Salesses, P., Schechtner, K., and Hidalgo, C.A., 2013. The collaborative image of the city: mapping the inequality of urban perception. *PloS One*, 8 (7), e68400.
Schroff, F., Kalenichenko, D., and Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.
Smith, E.E., and Medin, D.L., 1981. *Categories and concepts*. Cambridge, MA and London, England: Harvard University Press. Available from: [Accessed 20 Aug 2024].
Spruyt, V., 2018. Loc2vec: Learning location embeddings with triplet-loss networks. *Sentiance web article:* https://www. sentiance. com/2018/05/03/venue-mapping.
Su, H., *et al.*, 2015. Multi-view convolutional neural networks for 3d shape recognition. *In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 945–953.
Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46 (sup1), 234–240.
Tversky, A., 1977. Features of similarity. *Psychological Review*, 84 (4), 327–352.
Uber Technologies, Inc., 2024. H3: A hexagonal hierarchical spatial index. Available from: https://h3geo.org/ [Accessed 20 Aug 2023].
Vainio, T., *et al.*, 2019. 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 (CHI EA '19). New York, NY: Association for Computing Machinery, 1–8.

Wan, S., *et al.*, 2021. Spatial analysis and evaluation of medical resource allocation in china based on geographic big data. *BMC Health Services Research*, 21 (1), 1084.

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

Woźniak, S., and Szymański, P., 2021. Hex2vec: Context-aware embedding h3 hexagons with openstreetmap tags. *In: Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GEOAI '21)*. New York, NY: Association for Computing Machinery, 61–71.

Xi, Y., *et al.*, 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. *In: Proceedings of the ACM Web Conference 2022 (WWW '22)*. New York, NY: Association for Computing Machinery, 3308–3316.

Zeile, P., *et al.*, 2015. Urban emotions–tools of integrating people's perception into urban planning. *In: REAL CORP 2015*. 905–912.

Zhang, F., *et al.*, 2024a. Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. *Annals of the American Association of Geographers*, 114 (5), 876–897.

Zhang, F., *et al.*, 2018. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180, 148–160.

Zhang, J., *et al.*, 2020. 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 (CHI EA '20)*. New York, NY: Association for Computing Machinery, 1–8.

Zhang, M., and Bandara, A.K., 2024. Understanding pedestrians' perception of safety and safe mobility practices. *In: Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. New York, NY: Association for Computing Machinery, 1–17.

Zhang, Y., Li, Y., and Zhang, F., 2024b. Multi-level urban street representation with street-view imagery and hybrid semantic graph. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218, 19–32.