



Delft University of Technology

Computer vision-enriched discrete choice models, with an application to residential location choice

van Cranenburgh, Sander; Garrido-Valenzuela, Francisco

DOI

[10.1016/j.tra.2024.104300](https://doi.org/10.1016/j.tra.2024.104300)

Publication date

2025

Document Version

Final published version

Published in

Transportation Research Part A: Policy and Practice

Citation (APA)

van Cranenburgh, S., & Garrido-Valenzuela, F. (2025). Computer vision-enriched discrete choice models, with an application to residential location choice. *Transportation Research Part A: Policy and Practice*, 192, Article 104300. <https://doi.org/10.1016/j.tra.2024.104300>

Important note

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

Copyright

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.



Computer vision-enriched discrete choice models, with an application to residential location choice

Sander van Cranenburgh^{*}, Francisco Garrido-Valenzuela

CityAI Lab, Transport and Logistics Group, Delft University of Technology, The Netherlands

ARTICLE INFO

Keywords:

Discrete choice modelling
Computer vision
Residential location choice
Commute
Travel behaviour

ABSTRACT

Visual imagery is indispensable to many multi-attribute decision situations. Examples of such decision situations in travel behaviour research include residential location choices, vehicle choices, tourist destination choices, and various safety-related choices. However, current discrete choice models cannot handle image data algorithmically and thus cannot incorporate information embedded in images into their representations of choice behaviour. This gap between discrete choice models' capabilities and the real-world behaviour it seeks to model leads to incomplete and, possibly, misleading outcomes. To solve this gap, this study proposes "Computer Vision-enriched Discrete Choice Models" (CV-DCMs). CV-DCMs can handle choice tasks involving numeric attributes and images by integrating computer vision and traditional discrete choice models. Moreover, because CV-DCMs are grounded in random utility maximisation principles, they maintain the solid behavioural foundation of traditional discrete choice models. We demonstrate the proposed CV-DCM by applying it to data obtained through a novel stated choice experiment involving residential location choices. In this experiment, respondents faced choice tasks with trade-offs between commute time, monthly housing cost and street-level conditions, presented using images. We find that CV-DCMs can offer novel insights into preferences regarding features presented in images, such as what street-level conditions people find most and least attractive and how these preferences vary across age groups.

1. Introduction

Discrete Choice Models (DCMs) are widely used in transportation (and beyond) to describe how individual choices result from preferences over attributes and available alternatives in multi-attribute decision-making. When DCMs were invented in the 1970s, they were used to explain and predict mode and destination shares (McFadden, 1974; McFadden, 2001). Nowadays, DCMs are applied to a wide variety of choice situations, including residential location choice, route choice, vehicle choice, airport choice, time of day choice and many more (de Jong et al., 2003; Guevara and Ben-Akiva, 2006; Hess et al., 2007; Prato, 2009; Pinjari et al., 2011; Beck et al., 2013; Hess and Daly, 2014). DCMs are built on the notion that attributes have numeric values or can be converted into numeric values, e.g. in the case of a categorical level. In other words, the attributes that jointly make an alternative only involve numbers.

Visual imagery is crucial to many multi-attribute decision situations, in and beyond transportation. For example, visual information is indispensable to residential location choices. In today's digital age, it is hard to imagine searching for a house on a real estate website without access to images. Other examples of such decision situations in transportation include vehicle choices, tourist destination

^{*} Corresponding author.

E-mail address: s.vancranenburgh@tudelft.nl (S. van Cranenburgh).

choices, transport infrastructure design choices and choices related to safety, such as where to cross a street on foot and whether a route is safe enough to cycle. The widespread use of visual imagery, e.g. on websites like [Zillow.com](https://www.zillow.com) and in Stated Choice (SC) experiments, can be attributed to the fact that it is easier for people to perceive and process information presented through images than information presented in text or numbers (Pinker, 1990). In addition, visual imagery provides valuable details about the alternative, such as scale, texture, or quality, that are difficult to convey through textual descriptions or numbers (Childers et al., 1985). For instance, in a residential location choice context, visual characteristics of the (built) environment (henceforth referred to as street-level conditions) such as “safeness”, “openness”, “continuity”, and “common orientation” cannot be easily expressed in numbers but can effectively be communicated by images. The COVID-19 pandemic brought the importance of street-level conditions to the forefront, with millions of white-collar workers relocating to suburban areas with better street-level conditions during the pandemic-induced remote work shift (Economist, 2022; Lee and Huang, 2022). Therefore, to accurately represent choice behaviour in multi-attribute situations that involve visual imagery, it is necessary to have choice models capable of working with image data.

However, present-day DCMs cannot handle image data directly and, therefore, cannot incorporate information from images into their representations of choice behaviour. The inability to handle image data in DCMs creates a stark contrast between the behaviour it seeks to model, where images are widely used, and what DCMs can do. Even when researchers deliberately use images in SC experiments to visualise information that is challenging to convey in numbers, the information embedded in the images is scantily accounted for (Cherchi and Hensher, 2015; see Hevia-Koch and Ladenburg, 2019 for a thorough discussion). DCMs’ inability to handle image data leads to incomplete and potentially misleading outcomes.

As a solution, this study proposes “Computer Vision-enriched Discrete Choice Models” (henceforth abbreviated as CV-DCMs). These models can handle choice tasks involving both numeric attributes and an image. CV-DCMs are grounded in Random Utility Maximisation (RUM) principles (McFadden, 2000; Hess et al., 2018). Therefore, CV-DCMs maintain the solid behavioural foundation of traditional DCMs while expanding their application to include image data. We demonstrate the effectiveness of the proposed CV-DCMs by shedding light on the importance of street-level conditions to residential location choice behaviour relative to travel-related factors, such as travel time and travel cost. To do so, we have developed and administered a novel stated choice experiment involving trade-offs between commute travel time, monthly housing cost (both numeric attributes) and street-level conditions (presented using images).

The main contribution of this paper is methodological. It contributes to the growing body of literature in the travel behaviour field that seeks to integrate machine learning and DCMs (e.g. Iglesias et al., 2013; Hurtubia et al., 2015; Rossetti et al., 2019; Sifringer et al., 2020; Arkoudi et al., 2021; Ramírez et al., 2021; van Cranenburgh et al., 2021; Szép et al., 2023). More specifically, our study can best be positioned in two streams of the literature. The first stream of literature concerns studies that map human *perceptions* of the urban environment using a combination of street view images and machine learning (Naik et al., 2014; Dubey et al., 2016; Liu et al., 2017; Zhang et al., 2018; Wei, et al., 2022; Zhang et al., 2022; Zhang et al., 2024; and see Ito et al., 2024 for a review). In this stream of literature, models are trained on survey data in which respondents are typically presented with two street view images and asked to indicate which image looks safer/more vibrant/livelier/etc. After training, these models are commonly used to generate spatial maps showing where the urban environment is perceived as safe/vibrant/lively/etc. Our study also uses street view images and machine learning but deviates from this literature because it concerns *preferences*. Although perceptions and preferences are closely related concepts, they are not the same. Preferences are grounded in the theory of choice behaviour (Samuelson, 1938; Luce, 1959; Lancaster, 1966) and govern what people choose and how they make trade-offs. In contrast, perceptions are subjective interpretations of sensory stimuli, which may influence but do not necessarily determine individuals’ choices (Wade and Swanston, 2013).

The second stream of related literature concerns studies seeking to understand choice behaviour (and thus preferences) in the presence of visual stimuli (i.e. images) by first encoding the information from images into tabular form and then estimating traditional choice models. Encoding information from images can be done manually by the researcher (see e.g. Arriaza et al., 2004; Zhao et al., 2022) as well as algorithmically by computer vision algorithms. Manual encoding is labour-intensive and imposes a strong limitation on the number of images that can be utilised. Noteworthy, Patterson et al. (2017) circumvent this challenge using artificially created images that reflect specific attribute levels, such as dwelling type and space between buildings. Studies taking the algorithmic encoding approach typically use object detection and semantic segmentation models to extract information from images (e.g. Rossetti et al., 2019; Ramírez et al., 2021). Directly encoding information from images into tabular form offers a significant advantage; the modelling results (i.e. the preference parameters) are directly interpretable. However, this approach critically relies on prior knowledge of the factors influencing (choice) behaviour and the accuracy of the information extraction (either by human annotators or algorithmic object detection and segmentation models). The model proposed in this study does not rely on prior knowledge of the factors influencing the choice behaviour or the accuracy of object detection models. Also, it preserves the interpretability of the model’s parameters associated with the numeric attributes. However, because our model is trained end-to-end and its encoding is ‘hidden’, insights regarding preferences over images cannot be derived from scrutinising the model’s parameters.

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

The remaining part of this paper is organised as follows. Section 2 describes the proposed CV-DCMs. Section 3 discusses the stated choice data collection effort and reports the sample statistics, descriptive results and details on the training of the model. Section 4 contains the main results. Section 4.1 presents the results from the CV-DCMs and compares model fit and parameter estimates with those of traditional discrete choice models, which do not account for images. Section 4.2 shows what the CV-DCM has learned about what decision-makers find relevant for their residential location choices. It provides face validity to the modelling results. Section 4.3 demonstrates the merits of the CV-DCM by showing how CV-DCMs can be used to deepen understanding of residential location

preferences. Finally, [Section 5](#) draws conclusions and discusses limitations and directions for future research.

2. Methodology

This section presents the methodology. [Section 2.1](#) introduces relevant models and concepts from computer vision. [Section 2.2](#) proposes the modelling framework. [Section 2.3](#) briefly discusses implementation details and training.

2.1. Preliminary: Computer vision models and concepts

Computer Vision (CV) is concerned with extracting meaningful information from images, videos, and other forms of visual data. CV models typically detect scenes and objects in images ([Gu et al., 2018](#)). Nowadays, CV models are applied in a wide range of applications and numerous fields. In transport, CV models are essential for future autonomous vehicles to perceive and understand their environment; in healthcare, CV models are used in medical imaging to aid in diagnosing diseases and abnormalities; and, in retail, CV models are used to track customer movement in stores. As CV models grow and become more powerful, they can perform increasingly sophisticated visual tasks ([Sevilla et al., 2022](#)). The largest CV models currently in use contain over 1 billion weights ([Zhai et al., 2022](#)).

The building blocks of images are pixels. A pixel represents a single point in an image and contains information about its colour and brightness. Each pixel has a spatial location ($h \times w$) and a colour value. Most colour images nowadays use three colour channels: Red (R), Green (G), Blue (B), and 8 bits per colour channel (implying three 0–225 values), with which it is possible to create a wide range of colours and shades. Mathematically, images are usually represented as 3D tensors, which are multi-dimensional arrays of numerical values. Tensors enable easy processing and manipulation of images using various mathematical operations and algorithms, especially in combination with GPUs. The three dimensions of an image tensor typically correspond to the image's width, height, and colour channels. Thus, an RGB colour image with a resolution of 900×600 pixels can be represented as a 3D tensor with a shape of (900, 600, 3), where the first two dimensions correspond to the height and width of the image and the third dimension corresponds to the colour channels. An image tensor of a 900×600 RGB colour image contains 1.6 m data points.

CV models typically have two main components: a feature extractor and a classifier, see [Fig. 1](#). The feature extractor is generally a deep neural network that is trained to extract relevant features from images. The output of the feature extractor is the feature map, which is a lower-dimensional vector representation of the image and captures its salient features. In other words, the feature map contains (most of) the information of the image but is more compact in form. Usually, a feature map (a.k.a. embedding) is a flat array of floating points. Nowadays, a so-called transformer architecture is the mainstay choice as a feature extractor in the CV field ([Dosovitskiy et al., 2020](#)). In contrast to traditional convolution-based architectures, so-called Vision Transformers (ViT) rely on self-attention and multi-head attention mechanisms to learn spatial relationships between different parts of the image. In a ViT architecture, the input image is divided into a grid of non-overlapping patches, which are linearly embedded to produce a sequence of feature maps. These feature maps are then processed by a series of transformer encoder layers to learn spatial relationships between the different parts of the image. The classifier is a separate component which is trained to classify the input image based on the feature map. Typically, the classifier is a Multilayer Perceptron (MLP) with one or more fully connected layers, producing a probability distribution over the different output classes.

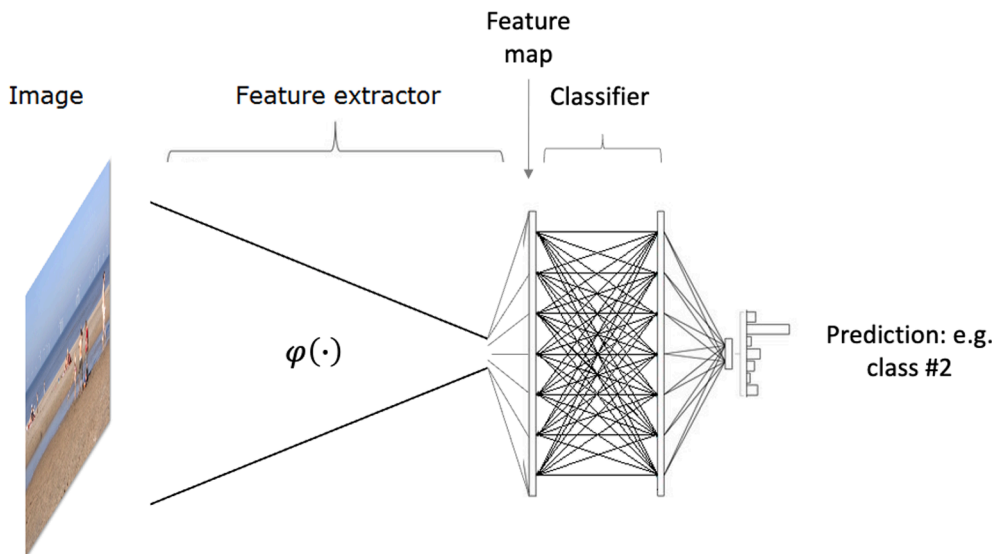


Fig. 1. Feature extraction and classification.

2.2. Computer vision-enriched discrete choice models

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

We assume decision-makers make decisions based on Random Utility Maximising (RUM) principles (McFadden 1974), see Equation (1), where U_{in} denotes the total indirect utility experienced by decision-maker n considering alternative i , V_{in} is the utility experienced by decision-maker n derived from attributes observable by the analyst. And, to account for the fact that the analyst does not observe everything that matters to the decision-maker's utility, an additive error term ε_{in} is added to each alternative (Train 2003).

$$U_{in} = V_{in} + \varepsilon_{in} \quad (1)$$

Furthermore, we assume decision-makers experience utility from both the numeric attributes X_i and the attributes encoded in the image \mathcal{I}_i , see Equation (2), where v is a preference function which maps the numeric attributes and the attributes encoded in the image of an alternative onto utility.

$$U_{in}(X_{in}, \mathcal{I}_{in}) = v(X_{in}, \mathcal{I}_{in}) + \varepsilon_{in} \quad (2)$$

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

1. We assume that the utility derived from the numeric attributes and the attributes encoded in the image are separable and additive in utility space, see Equation (3), where function f maps the (observed) numeric attributes onto utility and function g maps the attributes encoded in the image onto utility. Note that images typically encode multiple attributes. Therefore, the encoded attributes can be regarded as a composite good

$$U_{in}(X_{in}, \mathcal{I}_{in}) = f(X_{in}) + g(\mathcal{I}_{in}) + \varepsilon_{in} \quad (3)$$

2. We assume that utility is linear and additive with numeric attributes as well as with the attributes encoded in the images, as captured in the feature maps. Thus, f and g are standard linear-additive utility functions. As discussed in section 2.1, feature maps are more compact representations of images. Accordingly, we let $Z_i = \{z_{i1}, z_{i2}, \dots, z_{iK}\}$ denote the feature map of image \mathcal{I}_i , and $\varphi(\mathbf{w}) : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^K$ be a function that maps image \mathcal{I}_i onto feature map Z_i . Hence, φ is the transformation produced by the feature extractor of a CV model, and \mathbf{w} are its associated weights (i.e., the trainable parameters), which extracts the attributes encoded in the images. Both the numeric attributes X_i and feature map Z_i enter the utility function in a linear-additive fashion, as shown in Equation (4). In Equation (4), β_m denotes the marginal utility associated with attribute m ; x_{imn} denotes the attribute level of numeric attribute m of alternative i , as faced by decision-maker n ; and β_k denotes the weight associated with the k^{th} element of feature map Z_{in}

$$U_{in} = \underbrace{\sum_m \beta_m x_{imn}}_{\text{Systematic utility derived from numeric attributes}} + \underbrace{\sum_k \beta_k z_{ikn}}_{\text{Systematic utility derived from attributes encoded in the image}} + \varepsilon_{in} \quad (4)$$

where $Z_{in} = \varphi(\mathcal{I}_{in} | \mathbf{w})$

The reason that we let feature maps, as opposed to individual pixel values, enter the (indirect) utility function is that letting pixels enter the utility function is at odds with the notion that utility is derived from consuming a certain bundle of goods and services. After all, pixels are not consumed; rather, utility is derived consuming a good, which is conceptualised in terms of their constituent attributes (Lancaster, 1966). The feature map comprises the consumable attributes, encoded by the images' pixels. Thus, in the CV-DCM, $\varphi(\mathbf{w})$ produces a feature map containing the street-level attributes encoded in the image that are relevant to explain the choice behaviour (and in such a way that they map linearly onto utility). However, it should be noted that its elements do not come with any a priori behavioural or semantic interpretation. The semantic meaning of the element may be extracted through post-hoc eXplainable AI (XAI) analyses.

3. In line with common practice in choice modelling, we assume ε_{in} is independent and identically extreme Value type i distributed with a variance of $\pi^2/6$, resulting in the well-known and convenient closed-form logit formula for the choice probabilities (P_{in}), given in Equation (5), where C_n denotes the set of alternatives presented to decision maker n . Note that this assumption would, from a machine learning perspective, be equivalent to saying that the output layer is a Softmax function

$$P_{in} = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \quad (5)$$

Fig. 2 depicts a graphical representation of the model structure of the proposed CV-DCM. It shows that the network's upper and lower parts are identical. In the machine learning literature, this is referred to as a Siamese network (Bromley et al., 1994). This highlights an essential aspect of the CV-DCM's architecture: its consistency with RUM. It is consistent with RUM because it satisfies two conditions: regularity and transitivity (see Hess et al., 2018 for a rich discussion on RUM consistency and RUM consistency tests). This is evident from equation (4), which shows that (1) the utility of one alternative does not depend on the attributes of another alternative, and (2) the utility function preserves ordinality. As a result, we can conceive the values at nodes in the last layer as utilities. However, even though we can interpret the last layer as utility, we cannot interpret β_k in the same way we can with β_m . β_k can be conceived as a marginal utility – after all, it reflects the change in utility by a unit change in the attribute level. But, because the meaning and units of the elements on the feature map, Z_i , are unclear, they do not carry a behavioural meaning. Furthermore, although it is technically possible (though challenging) to compute standard errors associated with β_k , this is not immediately a meaningful thing to do. After all, without interpretation of elements of the feature map, we do not have a hypothesis we wish to accept or reject. Having said that, in the situation in which meanings are attributed to certain elements of the feature map – e.g. through the use of XAI – computation of the standard errors could become meaningful.

2.3. Feature extractor and training

In this study, we use the feature extractor of the DeiT base model (Touvron et al., 2021). DeiT models are data-efficient vision transformer-based models that produce competitive capabilities on benchmark data sets, such as ImageNet (Russakovsky et al., 2015), at a lower computational cost and data requirements than many of its competitors (Touvron et al., 2021). The DeiT base model comprises a relatively modest 86 million weights and produces feature maps containing $K = 1,000$ elements. Furthermore, we use transfer learning to train our CV-DCM (Bengio, 2012) to lower the computational time and amount of training data. The idea of transfer learning is to use a pre-trained network as the starting point for developing another network for a closely related task. In other words, rather than retraining the whole model from scratch, we start the training from an already good starting point when we train the CV-DCM. Our pre-trained DeiT base model is trained on ImageNet (Deng et al., 2009), a widely used benchmark image data set containing 1.2 million training images with 1,000 object classes.

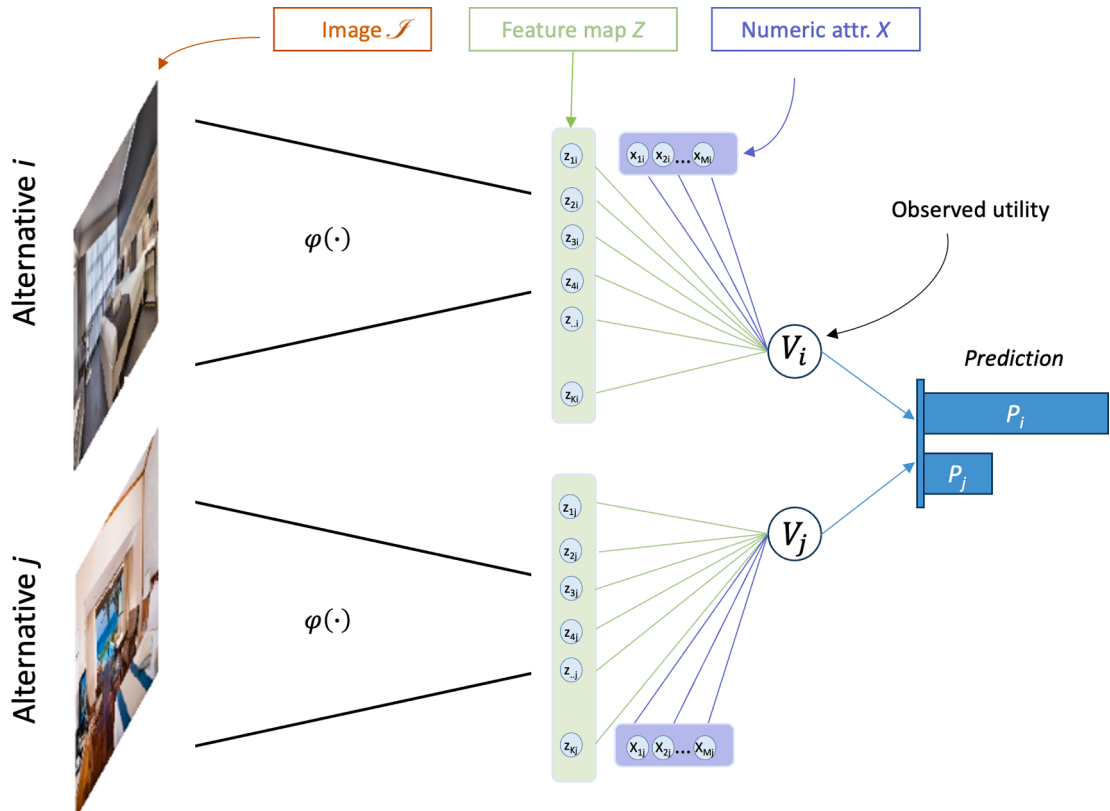


Fig. 2. Model structure of CV-DCM.

3. Data collection and training

We demonstrate the proposed CV-DCM by applying it to data obtained through a stated choice experiment involving residential location choices. The residential location choice makes a suitable case study because both numeric attributes and street-level conditions, which we visualise using images, can be expected to be important to residential location choice behaviour (Smith and Olaru 2013). Street-level conditions can be thought of as a composite good that encompasses various elements, such as cleanliness, greenery, infrastructure quality, and overall aesthetic appeal. Moreover, images of the sort that we need for conducting a residential location choice experiment, namely street-level images, are widely available from map services such as Google, Apple, and Baidu and have been used in numerous scientific inquiries, including research on safety perceptions and people's density in urban places (Dubey et al., 2016; Ito and Biljecki 2021; Ma et al., 2021; Garrido-Valenzuela et al., 2022). Having access to a sufficiently large and diverse set of images is crucial for effectively training the feature extractor of the CV-DCM. While the exact number of images required is unknown before training, more images (and choice observations) generally lead to better training. In addition to their availability, street-level images have been shown to be a reliable representation of street-level conditions, as demonstrated by Hanibuchi et al. (2019).

3.1. Stated choice experiment

In the Stated Choice (SC) experiment, we asked respondents to imagine they were required to move to a different neighbourhood. They were presented with two alternatives for residential locations and asked to indicate which of the two they would choose. Fig. 3 shows a screenshot of a choice task from the experiment. Prior to starting the choice experiment, respondents were provided with the following information:

1. Your new house is identical to your current house in terms of, e.g. size, type, built-year, furniture, maintenance, etc. Only your neighbourhood changes.
2. Your monthly housing cost (including rent, mortgage, taxes, insurance, etc.) may go up or go down.
3. Your new neighbourhood is relatively near your current neighbourhood, but your commute time may still go up or down. The commute time is for your current mode of transport.
4. Your situation stays the same in all other aspects, e.g. in terms of distances to amenities, schools, the general practitioner, etc.
5. The images shown in the choice tasks depict the window view at ground level on the street side.

The alternatives comprise two salient numeric attributes: monthly housing costs (*hhc*) and commute travel time (*tti*). We choose these two attributes for three reasons. Firstly, they are known to be important to the residential location choice (Tillema et al., 2010). Secondly, they apply generically to almost everyone's residential location choice. Thirdly, they may help to interpret our empirical results. The combination of cost and time attributes allows us to compute the Value-of-Travel-Time (VTT), a metric that is widely studied in transport (Small 2012) and thus can be used for model validation. Finally, we did not include more attributes to the design because the paper's objective is to demonstrate the effectiveness of the proposed CV-DCMs to capture visual preferences instead of, e.g. developing a comprehensive model to predict residential location choices.

As can be seen in Fig. 3, we have opted for a pivoted experimental design. We use a pivoted design to present respondents with as realistic choice situations as possible. Using absolute levels instead of pivoted levels would presumably render many choice tasks unrealistic because of the considerable variation across respondents' current situations, especially regarding housing costs. For the attribute housing cost, we have used seven pivoted levels. For the attribute travel time, the number of levels and ranges we presented to the respondent depended on the respondent's current travel time, see Table 1. The ranges of both attributes were determined through a small pilot conducted before the actual survey.

3.1.1. Street-level images

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

Importantly, for each image in our database, we also stored the month of the year in which the image was taken. The Netherlands lies in temperate zones, having four distinct seasons. Even though street-view images are usually collected on dry days, due to the seasonality, street-view images taken in the winter may look different from those taken in summer. These differences might, in turn, impact the utility experienced by the respondent from the depicted local environment (and thus must be accounted for in our models).

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

Which option would you choose?



	Option A	Option B
Your new street-view		
Monthly housing cost	€0 equally expensive as present	↑ €225 more expensive than presently
Commute travel time	↓ 5 minutes quicker than presently	↓ 10 minutes quicker than presently
	<input type="radio"/> Option A	<input type="radio"/> Option B

Fig. 3. Screenshot of the pivoted stated choice experiment.
(Image source: Google) (translated to English; original in Dutch)

Table 1
Attribute levels Stated Choice experiment.

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

3.1.2. Experimental design

We have used a random experimental design. Because the images do not possess ordinal or categorical levels, adopting an orthogonal or efficient experimental design strategy was not feasible, at least not considering the images. Therefore, we took a two-step approach to construct the choice tasks. First, we randomly pulled a pair of images from our image database. The only requirement imposed on the drawing was that the drawn images were not from the municipality where the respondent lives. We determined each respondent's municipality (and province) based on the postcode we elicited at the start of the survey. We excluded images from the respondents' municipalities to avoid unobserved heterogeneity entering our experiment, which may be derived from respondents' knowledge of places where the images were taken. Unobserved utilities flowing into stated choice experiments could lead to biased modelling outcomes if not econometrically accounted for (see, e.g., Train and Wilson 2008; Van Cranenburgh et al., 2014; Guevara and Hess 2019). While excluding images from respondents' own municipalities does not guarantee that respondents do not recognise the places the street-view images were taken, it lowers the probability.

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

consists of a trade-off between housing cost and travel time.

3.2. Data collection and sample description

The survey was implemented in SurveyEngine software and conducted in September 2022. The survey started with a few questions to determine respondents' eligibility for the survey. In particular, we elicited respondents' age, gender, postcode, and current commute travel time. Then came the SC experiment, in which each respondent was presented with 15 choice tasks. The images used in the choice tasks were directly retrieved from Google servers based on the urls from our image database. The survey ended with a series of questions regarding the respondents' current housing situation (e.g. housing costs, rating of the current visual street-level conditions) and commute situation (e.g. mode of transport, number of commute days). Noteworthy, we also asked respondents how important the three attributes (housing cost, travel time and street-level conditions) were for their decisions on a scale from 1 to 10. Although it is well-known that direct elicitation of preferences is treacherous (Nisbett and Wilson 1977), it still can provide first (albeit inconclusive) evidence of the importance of the street-level conditions, presented using the images, relative to the numeric attributes for the residential location choices.

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

Table 2 shows the sample statistics. Overall, the sample is representative of the target population. Also, for the variables that are not explicitly considered during the data collection, such as the modal split and household composition, the statistics are close to the population data (c.f. Ton et al., 2019). Furthermore, looking at the reported monthly housing cost, we notice that the largest share of the respondents has a housing cost below €750. This seems reasonable since the average net housing cost of rental houses in the Netherlands is around €700p/m; homeowners' average net housing cost is slightly above €900p/m (Stuart-Fox et al., 2022).

3.3. Descriptive analysis

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

Fig. 5 shows the Pearson correlation coefficients between importance ratings and a selection of respondent characteristics. Interestingly, the top row shows that the importance of the street-level conditions correlates strongest with the self-reported rating of respondents' current visual street-level conditions. This strong positive correlation suggests that people living in visually attractive neighbourhoods consider their visual street-level conditions relatively more important than people living in visually less attractive places. This observation aligns with Lee and Waddell (2010), who also find that the current situation affects residential location choice behaviour. Moreover, we see that the importance of the street-level conditions positively correlates with living in a detached or semi-detached house. A self-selection mechanism could explain this effect: people caring about their visual street-level conditions are more likely to choose an attractive residential location (see e.g., Van Wee 2009 for discussions on self-selection effects in residential location choices; Cao 2014). Finally, perhaps somewhat counter to expectations, we see that variables such as gender and monthly housing costs do not strongly correlate with the importance given to street-level conditions.

Furthermore, Fig. 5 reveals that the importance of the monthly housing cost (middle row) correlates strongest with living in house type 'Flat, gallery, porch, or apartment'. This correlation seems in line with intuition, given that low-income people are more likely to live in this type of housing. Finally, we see that the importance of the commute travel time (bottom row) positively correlates with age class 18–39 years. Since this age class sits in the centre of the working-age population, it makes sense that commute travel time is essential to this group. Altogether, the correlations reported in Fig. 5 seem plausible.

Next, we analyse the images used in the stated choice experiment. Although our street-view image database comprises urls to over 60k images, only slightly over 7.5k unique images are used in the stated choice experiment. Because images are drawn randomly from our image database with replacement, we expect that some images will be sampled more than once. Indeed, most images are used once. However, contrary to our design intentions, some images are used 20 times or more. A possible underlying cause could be the seed

¹ See <https://www.cint.com>.

Table 2
Sample statistics.

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

numbers used by the survey platform's software. Nevertheless, regardless of this issue's origin, when we deal with the issue carefully during the training of our models (see [Section 3.4.4](#)), it does not need to have an impact on our (substantive) findings.

Finally, [Fig. 6](#) shows the distribution of the month of the year of the images used in the survey. In line with expectations, the images are not evenly distributed over the year. We see that most images are taken in spring and summer (March to September). Furthermore, we notice that images have been sampled for all 12 months. This implies we can account for the impact of the seasons on the utility derived from the street-view images by estimating constants for all months (except one, which we need to fix to zero for normalisation).

How important were the for your choices?

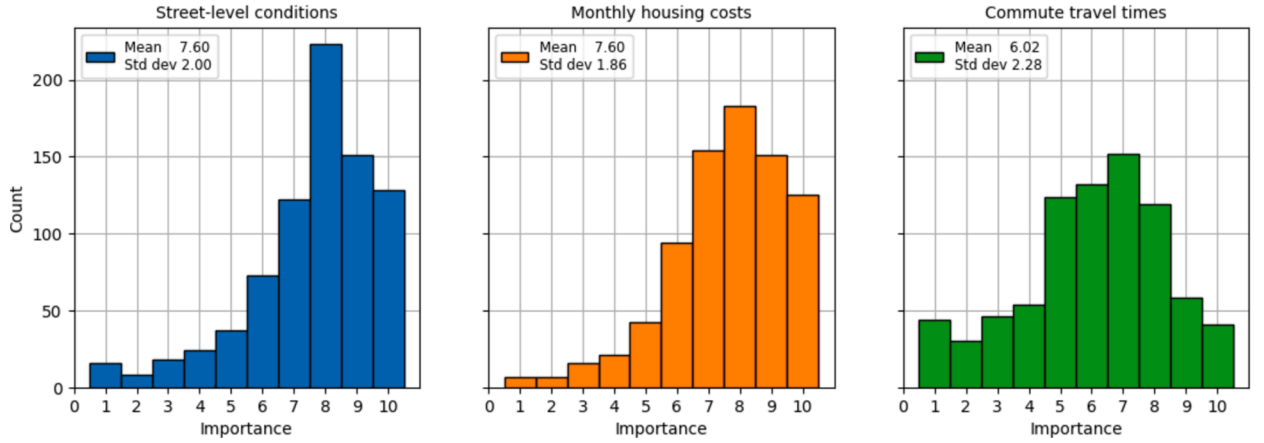


Fig. 4. Self-reported importance levels of attributes in the SC experiment.

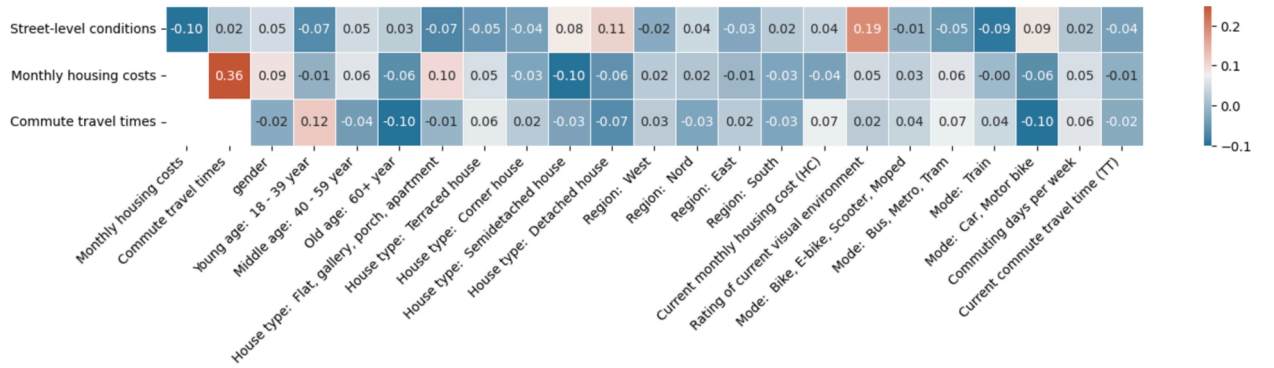


Fig. 5. Pearson correlation coefficients between importance ratings and respondent characteristics.

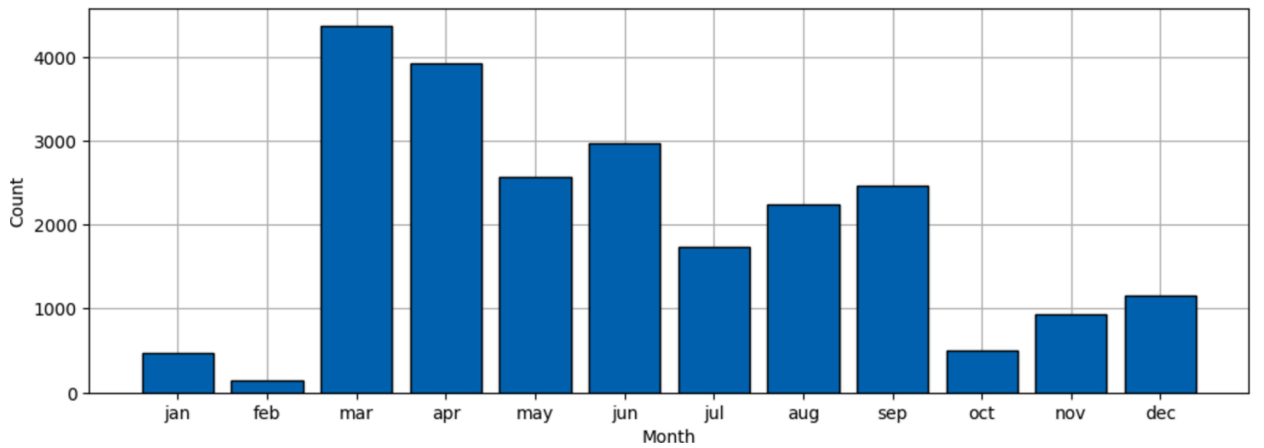


Fig. 6. Distribution of images used in the stated choice experiment over the months of the year.

3.4. Training

3.4.1. Loss function and implementation

Training a CV-DCM involves finding the weights of the model (β , w) that minimise the loss function. In other words, the weights of

the feature extractor and preference parameters of the utility function are jointly optimised. For this study, we use a cross-entropy loss function with an L2 regularisation term, see Equation (6). Minimising the cross-entropy loss is equivalent to maximising the Log-Likelihood of the data given the model – which is common practice in the choice modelling literature. The L2 regularisation aims to reduce the chance of model overfitting by penalising the magnitude of the weights in the model. γ governs the strength of the regularisation. Note that we apply regularisation only to w and not to preference parameters β_m and β_k . Regularising preference parameters could lead to undesirable biases.

$$\hat{w}, \hat{\beta} = \underset{w, \beta}{\operatorname{argmin}} \left[\underbrace{\frac{-1}{N} \sum_{n=1}^N \sum_{j=1}^J y_{nj} \log(P_{nj} | X_{nj}, \mathcal{J}_{nj}, \beta, w)}_{\text{cross-entropy loss}} + \underbrace{\gamma \sum_r w_r^2}_{\text{L2 regularisation}} \right] \quad (6)$$

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

3.4.2. Implementation and hyperparameter tuning

Our CV-DCM is implemented and trained in PyTorch (Paszke et al., 2019). PyTorch is a Python-based machine learning package commonly used for deep learning computer vision research because it supports GPU computing. We conducted hyperparameter tuning, in which we performed a grid-search over the most important hyperparameters: the optimisation algorithm, learning rate, batch size, and L2 regularisation (see Table 3). All other (hyper)parameters (such as dropout rates, layers and activation functions) were kept at their default values.

3.4.3. Image transformation and feature scaling

In line with common practice in computer vision, we transform and augment images while training the CV-DCM. Specifically, we conduct the two operations. First, we downsampled the images to 224×224 pixels. This downsizing operation ensures that images have the input dimensions expected by the CV model (i.e. DeiT base model). Second, we randomly flip images horizontally. This data augmentation operation reduces the model's ability to remember images, thus lowering the chance of overfitting the training data. Furthermore, we scale the numeric features. Scaling the features helps the optimiser to avoid getting stuck in local minima (Géron 2019). The most common type of scaling in machine learning involves shifting and scaling the features to a zero mean and a unit variance. We use another commonly used scaling technique to scale the housing cost and travel time features, called min–max scaling. This scaling entails scaling the features to a range of $[-1, 1]$. The advantage of this scaling technique is that it is straightforward and facilitates easy interpretation of the model's parameters. To facilitate interpretation, we have used the same scaling for all data (thus ignoring that the minimum travel time level varied across respondents, see Table 1). Specifically, all housing costs are divided by 225 and travel times are divided by 15.

3.4.4. Train-test split

Splitting the data into a train set and a test set is essential for training virtually all machine learning models because their high capacity makes them prone to overfitting (Géron 2019). As the name suggests, the train set is used for training the model; the test is unseen by the model during training and used to evaluate (test) the model's generalisation performance after training. If a trained model overfits the data, a gap in the performance between the train and test set will tell.

The most common way to create the train-test split is by randomly allocating observations to the two sets. When splitting data, it is important to avoid “data leakage”. Data leakage happens when the model has access to information during training that it does not have when deployed after training (see e.g. Hillel (2021) for its impact on choice model outcomes). For this study, we split our data across images. Thereby, we aim to avoid potential data leakage from learning the utility levels of specific images rather than generalisable high-level utility-generating features embedded in the images. Making such a split is, however, a nontrivial network problem. Every image is connected at least to one other image (the other street-view image presented in the choice task). However, some images are connected to dozens of other images because they are used more than once (see Section 3.3). Hence, when we assign one image to the train set, we must also place all directly and indirectly connected images in the train data set.

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

4. Results

We estimate/train four models on the residential location choice data whose utility functions are given in Equation 7 to Equation

² github.com/TUD-CityAI-Lab/Computer-vision-enriched-DCMs.

Table 3
Hyperparameter tuning CV-DCM.

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

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

$$\text{Model 1 } U_{in} = \beta_{hhc} hhc_{in} + \beta_{tii} tii_{in} + \varepsilon_{in} \quad (7)$$

$$\text{Model 2 } U_{in} = \beta_{hhc} hhc_{in} + \beta_{tii} tii_{in} + \sum_{mo} \beta_{mo} I_{in} + \varepsilon_{in} \quad (8)$$

$$\text{Model 3 } U_{in} = \beta_{hhc} hhc_{in} + \beta_{tii} tii_{in} + \sum_{mo} \beta_{mo} I_{in} + \sum_k \beta_k z_{ikn} + \varepsilon_{in} \quad (9)$$

$$\text{Model 4 } U_{in} = \beta_{hhc} hhc_{in} + \beta_{tii} tii_{in} + \sum_{mo} \beta_{mo} I_{in} + \sum_{age} \sum_k \beta_k^{age} \times age \times z_{ikn} + \varepsilon_{in} \quad (10)$$

Table 4
Estimation results.

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

^IUsing 4 CPUs (Intel Xeon @3.60 GHz).

^{II}Using GPU (GeForce RTX 2080Ti).

^{III}Obtained through computing the hessian while keeping the utility derived from the image fixed.

^{IV}Optimal hyperparameters: {optimiser: SGD, Batch size: 20, L2: 0.1, Learning rate: $1e^{-6}$ }.

^VOptimal hyperparameters: {optimiser: SGD, Batch size: 24, L2: 0.1, Learning rate: $1e^{-5}$ }.

$$I_{in} := \begin{cases} 1 & \text{if } mo = \mathcal{J}_{in}^{mo}, \\ 0 & \text{else} \end{cases}$$

$$Z_{in} = \varphi(\mathcal{J}_{in}|w)$$

$$\varepsilon_{in} \sim \text{i.i.d. Extreme Value Type I}$$

4.1. Estimation results

Table 4 shows the estimates for the behavioural interpretable parameters as well as the model performance on the train and test sets, using three (related) metrics: the Log-Likelihood, rho-square and cross-entropy. A good model performance on the test set implies the model generalises well to new/unseen data. It is important to note that we compare the model performance of the CV-DCMs with Models 1 and 2 to get a feeling of how much of the unexplained variance is captured by adding the computer vision model. Models 1 and 2 are not meant as a yardstick to show that the CV-DCMs outperform them. Because models 1 and 2 do not account for the images, they are unlikely to be used in practice on these data. In this regard, a more even-handed comparison would be models that first encode information in the images in tabular form, such as done by Ramírez et al. (2021). However, such a comparison would involve a study in itself and thus go beyond the scope of this paper.

We can draw three conclusions based on the performance metrics in Table 4. The first and most important conclusion is that the CV-DCM can extract relevant information from the street-level images to predict the choice behaviour. Looking at the generalisation performance, we see that the plain vanilla CV-DCM (Model 3) outperforms the two benchmark models (Model 1 and 2) by a fair margin. Specifically, the CV-DCM improves the Log-Likelihood on the test set by 57 Log-Likelihood points, and the rho-square jumps from 0.116 to 0.158.³ Since Model 3 collapses into Model 2 when setting $\beta_k = 0 \forall k$ (see Equations 8 and 9), we can statistically compare their model fit using a Likelihood Ratio Statistic (LRS). The LRS exceeds far the critical level of significance (set at $\alpha = 0.05$), with $K = 1,000$ degrees of freedom, supporting the notion that the CV-DCM's capability to handle images leads to a statistically significant improvement in model fit. Second, the month of the year carries limited information regarding the utility generated by the images, at least when used in isolation from other information from the images, as in Model 2. Comparing Models 1 and 2, we observe that Model 2 outperforms Model 1 by 23 Log-Likelihood points on the train set but performs on par on the test set. Hence, the incorporation of the month of the year in the utility function does not improve the generalisability of the conventional RUM-MNL models. Third, comparing Models 3 and 4, we see that allowing for an interaction between the feature map and age category (young, middle, old) further improves the model's generalisation performance. This reveals the presence of systematic taste heterogeneity concerning street view conditions across age groups.

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

Next, we look at the estimated taste parameters. We see that housing cost and commute travel time are highly relevant attributes to the residential location choice. In line with expectations, β_{tti} and β_{hhc} are highly significant, and their minus signs align with behavioural intuition. Based on β_{tti} and β_{hhc} , we also compute the VTT.⁴ In the context of our SC experiment, the VTT gives the (mean) willingness to pay per month for a one-hour travel time reduction per commute trip. A VTT between €217 and €228 per hour per month seems reasonable, considering that most respondents in our sample commute five days per week, and thus about 20 days per month. Furthermore, in line with expectations, we observe that the VTT is stable across all models. We expect stable $\beta_{tti} / \beta_{hhc}$ ratios because our experimental design is constructed in such a way that images and numeric attribute levels within choice tasks are entirely uncorrelated. Cramer (2005) shows that ratios of logit model estimates are unaffected by omitted variables if the omitted variables are uncorrelated with other explanatory variables.

The signs of the estimates associated with the months of the year are mostly intuitive. These estimates reflect the average utility difference between an image taken in that month and images taken in December (which we fixed to zero). In Model 2, we see that the estimates associated with the months of the year are mostly positive and significant for the spring and summer months. This can be explained by the notion that images taken in these months are more likely to look more attractive to live than images taken in winter,

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

⁴ The VTT is computed using $VTT = 60 \left(\frac{225}{15} \right) \frac{\beta_{tti}}{\beta_{hhc}}$. The factor $\left(\frac{225}{15} \right)$ comes from the fact that the attributes are scaled before training.

for instance, because the weather is better. However, the positive and significant estimate for January counters this line of argumentation and is hard to explain. In Models 3 and 4, the estimates associated with the months of the year do not carry the same interpretation as under Model 2. The utility derived from an image in Models 3 and 4 is the sum of the utility from the image's feature map and the estimate associated with the month of the year of the image. As a result, we cannot see the estimates associated with these two utility sources in isolation. One noteworthy observation concerning the estimates related to the months of the year in Models 3 and 4 is that fewer estimates are significant than in Model 2.⁵ This observation aligns with statistical expectations. Because feature maps already contain information about the weather conditions in the month of the year, explicitly adding the month to the model provides comparatively less information to explain the choice behaviour. For example, an image in which trees that have shed their leaves reveals the image is probably taken in winter. See [Siffringer and Alahi \(2023\)](#) for recent work on handling data congruency.

Lastly, we analyse the contributions to utility differences between the right and left-hand side alternatives derived from the images' feature maps. To do so, [Fig. 7](#) shows three kernel density plots for the plain vanilla CV-DCM (Model 3). The left-hand side plot shows the total utility difference as predicted by the trained CV-DCM; the middle plot shows the utility difference from the numeric attributes; and the right-hand side plot shows the utility difference from the attributes encoded in the images. We make several observations based on [Fig. 7](#). Firstly, looking at the range of x-axes of the middle and right-hand side plots, we see that the utility differences arising from the street-level conditions and numeric attributes are similar. This tells us that the part-worth utilities derived from the numeric attributes (housing cost and travel time) are of the same magnitude as those derived from the street-level conditions embedded in the street-view images.⁶ This observation adds to the evidence that street-level conditions are important to residential location choice behaviour and can effectively be modelled using images and CV-DCMs. Secondly, we notice that the distributions of utility differences are virtually equal for the test and train sets. This indicates that the CV-DCM does not overfit the training data, and the data are adequately split into train and test sets. Therefore, the CV-DCM must have learned to extract salient generalisable features from the images that generate utility. Thirdly, we see that the distribution of the utility differences stemming from the images is comparatively more bell-shaped than those of the numeric attributes. At first sight, this may seem odd, but it can be explained by how the choice tasks have been constructed. Recall that we removed choice tasks without trade-offs between the numeric attributes (see [Section 3.1.2](#) for more details). This removal leads to the bi-modal shape of the utility difference.

4.2. Face validity: What has the CV-DCM learned about street-level conditions?

The improvement in model performance by the CV-DCMs compared to the benchmark models supports the notion that the CV-DCMs can extract relevant information from images to predict choice behaviour. But, β_k and w do not carry a behavioural meaning. Therefore, they do not provide directly interpretable insights about what the CV-DCMs have learned regarding what decision-makers find important for their residential location choices. To shed light on what the CV-DCM has learned about the decision-makers' preferences, we show two collages of images taken from the test set, to which the trained CV-DCM (Model 3) assigns the highest ([Fig. 8](#)) and lowest ([Fig. 9](#)) utility levels. Note that the utility level is stamped in the top left of each image.⁷ These utility levels are 'uncorrected' for the month of the year. Hence, the top left image yields a utility of 1.63 if the image was taken in December, while it produces a utility of $1.63 - 0.12 = 1.51$ if it was taken in August (which it is).

What catches the eye in [Fig. 8](#) is that the images all look spacious, leafy and often water-abundant. We see many trees, grassland and detached houses. In the authors' view, these street-level conditions are indeed highly attractive. In sharp contrast, the images in [Fig. 9](#) look cramped, greyish, and urbanised and often have hallmarks of transportation, such as overhead wires, bus stops, parked bikes, and cars. In the authors' view, these street-level conditions are indeed highly unattractive. The mean difference in utility between the 20 best, shown in [Fig. 8](#), and the 20 worst street-level conditions, shown in [Fig. 9](#), is 2.7 utility points. The willingness to pay per month to move from the worst to the best street-level conditions can be computed by dividing the utility difference by β_{hhc} . The result yields a willingness to pay of 632 euros per month – which seems high but perhaps not implausible. Here, it should be noted that this estimate concerns the two most extreme street-level conditions.

4.3. Policy-relevant insights

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

4.3.1. Effect of age on preferences over street-level conditions

A policy-relevant insight that can be gleaned from the CV-DCM is the effect of age on preferences over street-level conditions. It is well established that different age generations have different housing needs (e.g. [Booi et al., 2021](#)). As such, it seems plausible that they

⁵ To compute the standard errors for Model 3 and Model 4, we fixed all weights w of φ and all β_k to their trained values, keeping the utility derived from the attributes encoded in the images constant. This approach is necessary, as constructing a Hessian matrix that accounts for all model parameters would be computationally infeasible. However, since potential covariances are not considered, the resulting standard errors may be less accurate.

⁶ Given the ranges of the numeric attributes presented in the SC experiment.

⁷ Note that the mean utility derived from the street-level conditions across all images is 0.02 (thus not precisely zero). This is inconsequential as utility has no absolute scale of level, only utility differences matter ([Train 2003](#)).

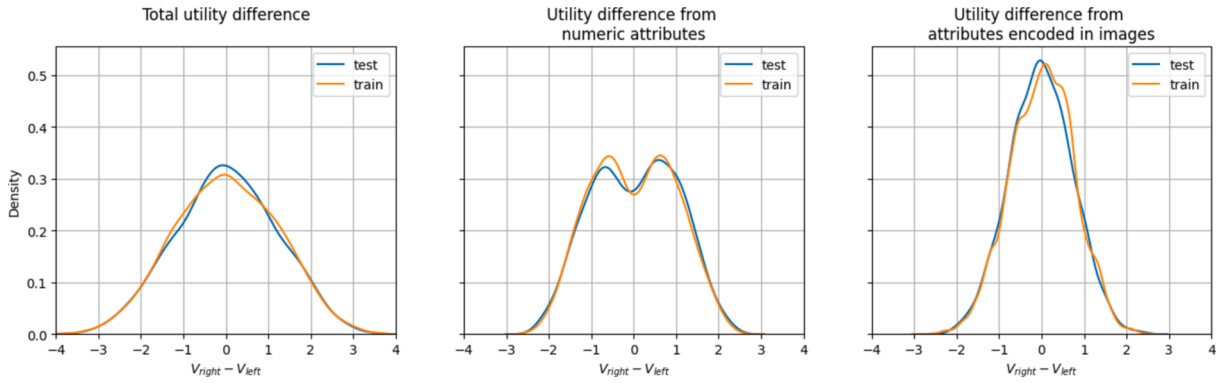


Fig. 7. Utility differences.



Fig. 8. Images showing street-level conditions with the highest predicted utility levels (based on the plain vanilla CV-DCM).

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

Figs. 10 and 11 show the results. More specifically, Fig. 10 shows sixteen street views that are comparatively more attractive to young people than older people; Fig. 11 shows sixteen street views that are comparatively more attractive to older people than young people. At the top left in each figure, the utility levels predicted by the CV-DCM for young and old people are shown. Furthermore, a kernel density plot shows the part-worth utility distribution (top) on the right-hand side of each figure. The vertical lines in this plot indicate where the sixteen depicted images sit in the overall distribution for younger (blue) and older (orange) people. A scatter plot scatters the part-worth utilities to young (x-axis) and old (y-axis) people at the bottom of the right-hand side plot. The red dots in the scatter plots correspond to the depicted street views.

Based on Figs. 10 and 11, a couple of policy observations can be made. Firstly, the scatter plots show a moderate correlation between the part-worth utilities from street-level conditions to younger and older people ($\rho = 0.76$). This means that, across the board, young and old people tend to agree on what attractive and unattractive street-level conditions are. But there are also street-level conditions where the utilities clearly diverge. In particular, Fig. 10 shows that young people find suburban areas relatively more attractive than old people. Most images in Fig. 11 show hallmarks of suburbia, like terraced houses, parking facilities, gardens,



Fig. 9. Images showing street-level conditions with the lowest predicted utility levels (based on the plain vanilla CV-DCM).



Fig. 10. Images showing street-level conditions that are comparatively attractive to younger people as compared to older people.

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

4.3.2. Relationship between visual attractiveness and population density

Faced with scarcity of land and increasing population levels, various Western European governments have developed housing policies with the aim of creating compact, high-density cities (e.g., by building more high-rises). Previous research, however, suggests that low-density (rural) areas are considered to be more visually appealing and scenic (Bijker and Haartsen, 2012) and that this



Fig. 11. Images showing street-level conditions that are comparatively attractive to older people as compared to younger people.

heightened visual attractiveness is one of the main motivations for “counter-urbanism” (Elshof et al., 2017), which is characterised by people moving away from urban areas and settling in rural or suburban areas. Counter-urbanising could thus undermine policies designed to create compact cities and put more strain on already burdened transportation networks. Previous studies into counter-urbanism have mostly relied on proxies for visual attractiveness, such as shares of older housing, proximity to natural areas, and the number of nearby hotels.

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

Fig. 12 presents the results of this analysis in a box plot. In line with the motivation for counter-urbanism, we find evidence that low-density (rural) areas have more attractive street-level conditions. In fact, using t-tests, we have established that the mean values of each population density quantile are significantly distinct from the means of all the other quantiles. This implies that there are notable differences in the utility of street-level conditions across levels of population density. From a policy perspective, these results suggest that policies aimed at creating compact cities should seriously take the attractiveness of street-level conditions into account. Although this study does not investigate counter-urbanism, failing to do so may undermine the effectiveness of such policies, as they may push people towards the suburbs and beyond.

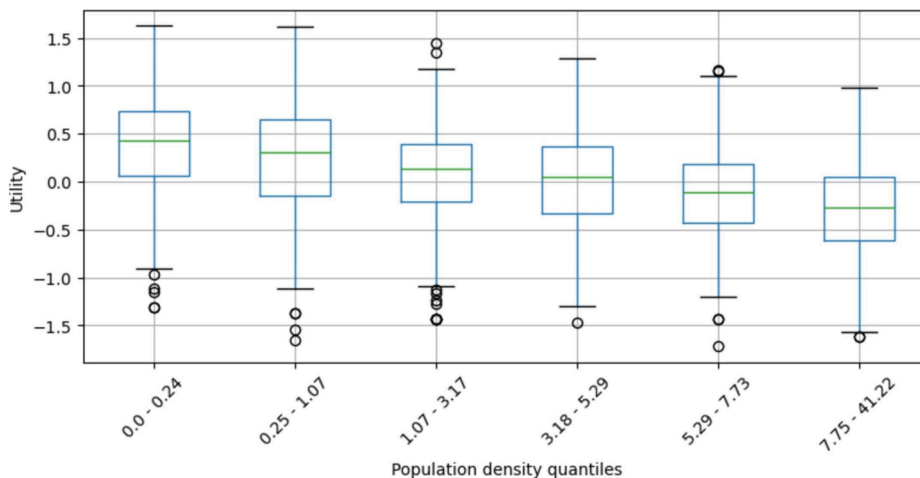


Fig. 12. Utility of street-level conditions as a function of population-density quantiles (based on Model 3).

5. Conclusion and discussion

This paper contributes to the recent methodological progress made in the fields of transportation and choice modelling that aims to bring machine learning and DCMs closer together (e.g. Sifringer et al., 2020; Arkoudi et al., 2021; Ramírez et al., 2021; van Cranenburgh et al., 2021). We have proposed a new choice model, called “Computer Vision-enriched Discrete Choice Models” (abbreviated as CV-DCM), for modelling multi-attribute choice behaviour in the presence of visual and numeric stimuli – methodologically expanding the realm of discrete choice models. The CV-DCM is built from behavioural assumptions, starting with random utility maximisation principles. As such, it has a solid behavioural foundation and can be used to derive marginal utilities and (in principle) willingness to pay estimates. The model should thus be conceived as a behaviour-informed choice rather than a behaviour-agnostic machine learning model. We have demonstrated its merits by applying it to residential location behaviour –which is strongly coupled with travel demand. We have shown that CV-DCMs can produce new insights into preferences over visual street-level conditions. Notably, we have uncovered which residential places people find most and least attractive and how attractiveness varies with population density.

The proposed model, in conjunction with SC experiments, can potentially enhance the understanding of other transport-related preferences in travel behaviour research. Using images can be particularly beneficial when numbers or text fail to convey the choice situation effectively. For instance, preferences related to crowdedness, traffic safety, and spaciousness may be better understood through the use of images in SC experiment showing, e.g. the crowdedness of train platforms, the safety situation of pedestrian crossings, or the extra legroom available when upgrading from economy to business class when booking flights. Incorporating such visualisations can provide valuable insights for transport planners and policymakers seeking to improve transportation systems and services. For instance, inspired by studies like Rossetti et al. (2019), in a follow-up study, we used our trained CV-DCM to assess the spatial distribution of utility derived from street-level conditions in residential location choices on a city-wide scale (van Cranenburgh and Garrido-Valenzuela, 2024).

This study raises a plethora of questions and opens up a multitude of avenues for further investigation at the intersection of choice modelling, computer vision and cognitive psychology. Two technical questions are of particular interest: (1) how to handle multiple images and (2) how to extract more and better information from trained CV-DCMs. Regarding the first question, multiple images are often used to describe alternatives. For instance, real estate websites like Zillow.com and online retailers like Amazon.com often use dozens of images per home or product. While the proposed modelling framework can accommodate a single image per alternative, future research can extend it to enable multiple images (e.g. inspired by Baevski et al., 2022). This methodological advancement would further expand the application domain and enhance the behaviour realism of the discrete choice models. The second question concerns how to extract more and better information from (trained) CV-DCMs. Recent developments in eXplainable AI (XAI) (Arrieta et al., 2020) offer a range of techniques that can be adapted to extract information from trained CV-DCMs. In particular, they can be leveraged to shed light on what features are learned by the model to explain the choice behaviour and help validate CV-DCMs. Such insights are potentially helpful not only for researchers but also for policymakers and urban planners. For example, in the context of the study’s application, XAI techniques should be able to provide insights into the features that make neighbourhoods attractive (as shown in Fig. 8) or unattractive (as shown in Fig. 9), which can inform the development of planning policies. Additionally, at present, the computation of the standard errors associated with the elements in the feature map (i.e. β_k) is technically challenging, inter alia, because of large covariance matrices. Future research could explore using smaller feature maps, achieved through techniques like pruning or semantic regularisation (Liao et al., 2016), as potential solutions to address this issue.

We conclude with a word of caution regarding the use of images in choice experiments. Although images hold great potential due to humans’ ability to extract information from them effectively, their incorporation into stated choice experiments must be approached cautiously. There are still many uncertainties surrounding their usage. For instance, using images could potentially skew attention to the images (and thus away from the numeric attributes). Its use might thus lead to underestimation of the estimates linked to the numeric attributes. Jansen et al. (2009) find some evidence supporting this observation, derived from a (small) survey wherein identical choice tasks were presented with and without accompanying impression photos. In connection with this, there is a risk of biased estimates associated with numeric attributes when CV-DCMs are trained on stated choice data wherein congruence exists between information in the image and numeric attributes (see recent work by Sifringer and Alahi, 2023). Hence, care must be taken that the information presented in images does not contain cues about the levels of the numeric attributes when designing stated choice experiments containing images. Another concern regarding the use of images is that people’s wishes and preferences may influence their visual perceptions (Balcetis and Dunning, 2006). Simply put, people may see what they want to see. This notion that images can be interpreted in multiple ways is also neatly illustrated by the iconic modern art painting “Ceci n’est pas une pipe”. The painting depicts a pipe. However, the artist of the painting, René Magritte, claims that it is not a pipe but a painting (readers interested in a more profound discussion of the painting are referred to Foucault, 1983). This highlights the challenge and need to align respondents’ interpretation of the images with the researcher’s intentions. Keys to the effective use of images in stated choice experiments can likely be found in the cognitive psychology field, which is concerned with studying mental processes such as perception, attention, and memory. Their insights can help researchers in our field to understand better how humans perceive and interpret visual information, which, in turn, can guide, e.g. what sort of images to use, how to present images (e.g. in relation to numeric attributes), and how to design SC experiments involving images more generally. In sum, further research is needed to comprehensively understand the implications and best practices regarding using images in choice experiments.

Finally, to fully harness the complementary information provided by text and images and pursue the avenues for future research outlined above, it is important to note that our modelling tools need a significant push. The current estimation software, survey platforms, computational resources and data handling practices in our field are not geared towards working with (large numbers of)

images. Moreover, working with a large number of images generally places higher technical demands on the programming and data-handling skills of researchers. Fortunately, these hurdles are surmountable. Open science practices and actively seeking cross-fertilisation between travel behaviour research, choice modelling, computer vision and cognitive psychology can accelerate progress. By sharing our data openly, we hope to contribute to this advancement.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT, a large language model developed by OpenAI, in order to refine the language and improve sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Sander van Cranenburgh: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Investigation, Formal analysis, Data curation, Conceptualization. **Francisco Garrido-Valenzuela:** Writing – review & editing, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by the TU Delft AI Labs programme. Additionally, we would like to thank the three anonymous reviewers for their constructive input during the development of this paper.

References

- Arkoudi, I., Azevedo, C. L., and Pereira, F. C. (2021). Combining Discrete Choice Models and Neural Networks through Embeddings: Formulation, Interpretability and Performance. *arXiv preprint arXiv:2109.12042*.
- Arriaza, M., Cañas-Ortega, J.F., Cañas-Madueño, J.A., Ruiz-Aviles, P., 2004. Assessing the visual quality of rural landscapes. *Landsc. Urban Plan.* 69 (1), 115–125.
- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58, 82–115.
- Baeviski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M., 2022, June. Data2vec: A general framework for self-supervised learning in speech, vision and language. In: *International Conference on Machine Learning*. PMLR, pp. 1298–1312.
- Balcetis, E., Dunning, D., 2006. See what you want to see: motivational influences on visual perception. *J. Personal. Soc. Psychol.* 91 (4), 612.
- Beck, M.J., Rose, J.M., Hensher, D.A., 2013. Environmental attitudes and emissions charging: An example of policy implications for vehicle choice. *Transp. Res. A Policy Pract.* 50, 171–182.
- Bengio, Y., 2012. Deep learning of representations for unsupervised and transfer learning. In: *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, pp. 17–36.
- Bijker, R.A., Haartsen, T., 2012. More than counter-urbanisation: Migration to popular and less-popular rural areas in the Netherlands. *Population, Space and Place* 18 (5), 643–657.
- Booi, H., Boterman, W.R., Musterd, S., 2021. Staying in the city or moving to the suburbs? Unravelling the moving behaviour of young families in the four big cities in the Netherlands. *Popul. Space Place* 27 (3), e2398.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1994. Signature verification using a “siamese” time delay neural network. *Adv. Neural Inf. Proces. Syst.*
- Cao, X., 2014. Examining the impacts of neighborhood design and residential self-selection on active travel: a methodological assessment. *Urban Geogr.* 1–20.
- Cherchi, E., Hensher, D.A., 2015. Workshop synthesis: Stated preference surveys and experimental design, an audit of the journey so far and future research perspectives. *Transp. Res. Procedia* 11, 154–164.
- Childers, T.L., Houston, M.J., Heckler, S.E., 1985. Measurement of Individual Differences in Visual Versus Verbal Information Processing. *J. Consum. Res.* 12 (2), 125–134.
- Cramer, J.S., 2005. Omitted variables and misspecified disturbances in the logit model. *Tinbergen Institute Discussion Paper*, Manuscript.
- de Jong, G., Daly, A., Pieters, M., Vellay, C., Bradley, M., Hofman, F., 2003. A model for time of day and mode choice using error components logit. *Transport. Res. Part E: Logistics Transport. Rev.* 39 (3), 245–268.
- Deng, J., Dong, W., Socher, R., Li, L., Kai, L., Li, F.-F., 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G. & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A., 2016. *Deep Learning the City: Quantifying Urban Perception at a Global Scale*. Springer International Publishing, Cham.
- Economist, T., 2022. How the pandemic has changed American homebuyers' preferences. *The Economist*, UK.
- Elshof, H., Haartsen, T., Van Wissen, L.J., Mulder, C.H., 2017. The influence of village attractiveness on flows of movers in a declining rural region. *J. Rural Stud.* 56, 39–52. *ISO* 690.
- Foucault, M., 1983. *This is not a pipe*. Univ of California Press).
- Garrido-Valenzuela, F., van Cranenburgh, S., Cats, O., 2022. Enriching geospatial data with computer vision to identify urban environment determinants of social interactions. *AGILE: Giscience Series* 3, 72.
- Géron, A., 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T., 2018. Recent advances in convolutional neural networks. *Pattern Recogn.* 77, 354–377.
- Guevara, C., Ben-Akiva, M., 2006. Endogeneity in Residential Location Choice Models. *Transp. Res. Rec.: J. Transp. Res. Board* 1977 (-1), 60–66.

- Guevara, C.A., Hess, S., 2019. A control-function approach to correct for endogeneity in discrete choice models estimated on SP-off-RP data and contrasts with an earlier FIML approach by Train & Wilson. *Transp. Res. B Methodol.* 123, 224–239.
- Hanibuchi, T., Nakaya, T., Inoue, S., 2019. Virtual audits of streetscapes by crowdworkers. *Health Place* 59, 102203.
- Hess, S., Daly, A., 2014. Handbook of choice modelling. Edward Elgar Publishing.
- Hess, S., Daly, A., Batley, R., 2018. Revisiting consistency with random utility maximisation: theory and implications for practical work. *Theor. Decis.* 84 (2), 181–204.
- Hess, S., Adler, T., Polak, J.W., 2007. Modelling airport and airline choice behaviour with the use of stated preference survey data. *Transport. Res. Part E: Logistics Transport. Rev.* 43 (3), 221–233.
- Hevia-Koch, P., Ladenburg, J., 2019. Where should wind energy be located? A review of preferences and visualisation approaches for wind turbine locations. *Energy Res. Soc. Sci.* 53, 23–33.
- Hillel, T. (2021). New perspectives on the performance of machine learning classifiers for mode choice prediction: An experimental review. 21st Swiss Transport Research Conference, Monte Verita, Ascona. URL: <http://www.strc>.
- Hurtubia, R., Guevara, A., Donoso, P., 2015. Using Images to Measure Qualitative Attributes of Public Spaces through SP Surveys. *Transp. Res. Procedia* 11, 460–474.
- Iglesias, P., Greene, M., Ortúzar, J.d.D., 2013. On the perception of safety in low income neighbourhoods: using digital images in a stated choice experiment. *The State of the Art and the State of Practice, Choice Modelling*, pp. 193–210.
- Ito, K., Biljecki, F., 2021. Assessing bikeability with street view imagery and computer vision. *Transport. Res. Part C: Emerging Technol.* 132, 103371.
- Ito, K., Kang, Y., Zhang, Y., Zhang, F., Biljecki, F., 2024. Understanding urban perception with visual data: A systematic review. *Cities* 152, 105169.
- Jansen, S., Boumeester, H., Coolen, H., Goetgeluk, R., Molin, E., 2009. The impact of including images in a conjoint measurement task: evidence from two small-scale studies. *Journal of housing and the built environment* 24, 271–297.
- Lancaster, K.J., 1966. A new approach to consumer theory. *Journal of political economy* 74 (2), 132–157.
- Lee, J., Huang, Y., 2022. Covid-19 impact on US housing markets: evidence from spatial regression models. *Spat. Econ. Anal.* 17 (3), 395–415.
- Lee, B.H.Y., Waddell, P., 2010. Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation* 37 (4), 587–601.
- Liao, Y., Kodagoda, S., Wang, Y., Shi, L., Liu, Y., 2016, May. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. In: 2016 IEEE international conference on robotics and automation (ICRA). IEEE, pp. 2318–2325.
- Liu, L., Silva, E.A., Wu, C., Wang, H., 2017. A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Comput. Environ. Urban Syst.* 65, 113–125.
- Luce, R.D., 1959. Individual choice behavior, 4. Wiley, New York.
- Ma, X., Ma, C., Wu, C., Xi, Y., Yang, R., Peng, N., Zhang, C., Ren, F., 2021. Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing. *Cities* 110, 103086.
- McFadden, D., 1974. The measurement of urban travel demand. Institute of Urban & Regional Development, University of California, Berkeley.
- McFadden, D., 2000. Disaggregate Behavioral Travel Demand's RUM Side, A 30 year retrospective. The leading, In *Travel behavior Research*, pp. 17–64.
- McFadden, D.L., 2001. Economic Choices. *Am. Econ. Rev.* 91 (3), 351–378.
- Naik, N., Philipoom, J., Raskar, R., Hidalgo, C., 2014. Streetscore-predicting the perceived safety of one million streetscapes. In: *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 779–785.
- Nisbett, R.E., Wilson, T.D., 1977. Telling More Than We Can Know - Verbal Reports on Mental Processes. *Psychol. Rev.* 84 (3), 231–259.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 2019. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inform. Process. Syst.* 32.
- Patterson, Z., Darbani, J.M., Rezaei, A., Zacharias, J., Yazdizadeh, A., 2017. Comparing text-only and virtual reality discrete choice experiments of neighbourhood choice. *Landsc. Urban Plan.* 157, 63–74.
- Pinjari, A., Pendyala, R., Bhat, C., Waddell, P., 2011. Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation* 38 (6), 933–958.
- Pinker, S., 1990. A theory of graph comprehension. *Artificial Intellig. Future Test.* 73, 126.
- Prato, C.G., 2009. Route choice modeling: past, present and future research directions. *Journal of Choice Modelling* 2 (1), 65–100.
- Ramírez, T., Hurtubia, R., Lobel, H., Rossetti, T., 2021. Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety. *Landsc. Urban Plan.* 208, 104002.
- Rossetti, T., Lobel, H., Rocco, V., Hurtubia, R., 2019. Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. *Landsc. Urban Plan.* 181, 169–178.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252.
- Samuelson, P.A., 1938. A note on the pure theory of consumer's behaviour. *Economica* 5 (17), 61–71.
- Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., Villalobos, P., 2022. Compute trends across three eras of machine learning. 2022 International Joint Conference on Neural Networks (IJCNN), IEEE.
- Siffringer, B., & Alahi, A. (2023). Images in Discrete Choice Modeling: Addressing Data Isomorphism in Multi-Modality Inputs. *arXiv preprint arXiv:2312.14724*.
- Siffringer, B., Lurkin, V., Alahi, A., 2020. Enhancing discrete choice models with representation learning. *Transp. Res. B Methodol.* 140, 236–261.
- Small, K.A., 2012. Valuation of travel time. *Econ. Transp.* 1 (1), 2–14.
- Smith, B., Olaru, D., 2013. Lifecycle stages and residential location choice in the presence of latent preference heterogeneity. *Environ Plan A* 45 (10), 2495–2514.
- Stuart-Fox, M., Klempner, T., Ligthart, D., Blijie, B., 2022. Wonen langs de meetlat. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, DEN HAAG.
- Szép, T., van Cranenburgh, S., Chorus, C., 2023. Moral rhetoric in discrete choice models: a Natural Language Processing approach. *Qual. Quant.*
- Tillema, T., Van Wee, B., Ettema, D., 2010. The influence of (toll-related) travel costs in residential location decisions of households: A stated choice approach. *Transp. Res. A Policy Pract.* 44 (10), 785–796.
- Ton, D., Duives, D.C., Cats, O., Hoogendoorn-Lanser, S., Hoogendoorn, S.P., 2019. Cycling or walking? Determinants of mode choice in the Netherlands. *Transp. Res. A Policy Pract.* 123, 7–23.
- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning, PMLR.
- Train, K.E., 2003. Discrete choice methods with simulation. Cambridge University Press, New York.
- Train, K., Wilson, W.W., 2008. Estimation of stated-preference experiments constructed from revealed-preference choices. *Transp. Res. B Methodol.* 42 (3), 191–203.
- Van Cranenburgh, S., Chorus, C.G., Van Wee, B., 2014. Vacation behaviour under high travel cost conditions – A stated preference of revealed preference approach. *Tour. Manag.* 43, 105–118.
- van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., Walker, J., 2021. Choice modelling in the age of machine learning-discussion paper. *J. Choice Model.* 100340.
- van Cranenburgh, S., & Garrido-Valenzuela, F. (2024). A utility-based spatial analysis of residential street-level conditions; A case study of Rotterdam. *arXiv preprint arXiv:2410.17880*.
- Van Wee, B., 2009. Self-Selection: A Key to a Better Understanding of Location Choices, Travel Behaviour and Transport Externalities? *Transp. Rev.* 29 (3), 279–292.
- Wade, N., Swanson, M., 2013. Visual perception: An introduction. Psychology Press.
- Wei, J., Yue, W., Li, M., Gao, J., 2022. Mapping human perception of urban landscape from street-view images: A deep-learning approach. *Int. J. Appl. Earth Observat. Geoinformation* 112, 102886.
- Zhai, X., Kolesnikov, A., Houslsby, N., Beyer, L., 2022. Scaling vision transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Zhang, A., Song, L., Zhang, F., 2022. Perception of pleasure in the urban running environment with street view images and running routes. *J. Geog. Sci.* 32 (12), 2624–2640.

- Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H., Ratti, C., 2018. Measuring human perceptions of a large-scale urban region using machine learning. *Landsc. Urban Plan.* 180, 148–160.
- Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Ratti, C., 2024. Urban Visual Intelligence: Studying Cities with Artificial Intelligence and Street-Level Imagery. *Ann. Am. Assoc. Geogr.* 114 (5), 876–897.
- Zhao, Y., van den Berg, P.E., Ossokina, I.V., Arentze, T.A., 2022. Comparing self-navigation and video mode in a choice experiment to measure public space preferences. *Comput. Environ. Urban Syst.* 95, 101828.