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

Sander van Cranenburgh[1]

Francisco Garrido Valenzuela[1]

[1]CityAI lab, Transport and Logistics Group, Delft University of Technology

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 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. As such, this research contributes to the growing body of literature in the travel behaviour field that seeks to integrate discrete choice modelling and machine learning.

1. Introduction

Discrete Choice Models (DCMs) are widely used in transportation (and beyond) to describe how individual choices result from preferences over attributes and available alternatives in multi-attribute decision-making. When DCMs were incepted in the 1970s, they were used to explain and predict mode and destination shares (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 (Brownstone et al. 2000; de Jong et al. 2003; Guevara and Ben-Akiva 2006; Hess et al. 2007; Prato 2009; Pinjari et al. 2011; 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. 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 travel behaviour research include vehicle choices, tourist destination choices, transport infrastructure design choices and choices related to safety, such as where to cross a street on foot and whether a route is safe enough to cycle. The widespread use of visual imagery, e.g. on websites like Zillow.com and in Stated Choice (SC) experiments, can be attributed to the fact that it is easier for people to perceive and process 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, street-level conditions such as "safeness", "openness", and "greenness" 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 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 scantly 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. To do so, we have developed and administered a novel stated choice experiment involving trade-offs between commute time (numeric attributes), monthly housing cost (numeric attributes) and street-level conditions (image).

Methodologically, this study contributes to the growing body of literature in the travel behaviour field that seeks to integrate machine learning and DCMs (e.g. 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). Our model shares similarities with Arkoudi et al. (2021) in its use of embeddings, but it extends this work by incorporating images. We also build on Ramírez et al. (2021) approach, which also seeks to explain choice behaviour in the presence of images. But, importantly, our model does not rely on object detection models, which can detect only a finite and predefined number of objects. Furthermore, this research also provides a significant contribution to the literature on residential location choice behaviour by highlighting the importance of street-level conditions in residential location choices.

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.

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 x 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. 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 x 600 pixels can be represented as a 3D tensor with a shape of (900, 600, 3), where the first two dimensions correspond to the height and width of the image and the third dimension corresponds to the colour channels. An image tensor of a 900 x 600 RGB colour image contains 1.6m data points.

CV models typically have two main components: a feature extractor and a classifier. The feature extractor generally is a deep neural network which 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. The classifier is a separate component which is trained to classify the input image based on the feature map extracted by the feature extractor. Typically, the classifier is a Multilayer Perceptron (MLP) with one or more fully connected layers, producing a probability distribution over the different output classes.

Traditionally, computer vision models have used Convolution Neural Networks (CNNs) architectures; see Figure 1. A CNN comprises a series of so-called convolution layers, which reduce the dimensionality of the data by taking a series of convolutions followed by an MLP classifier. More recently, transformer-based architectures have taken over the CV field (Dosovitskiy et al. 2020). In contrast to a CNN-based architecture, 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.

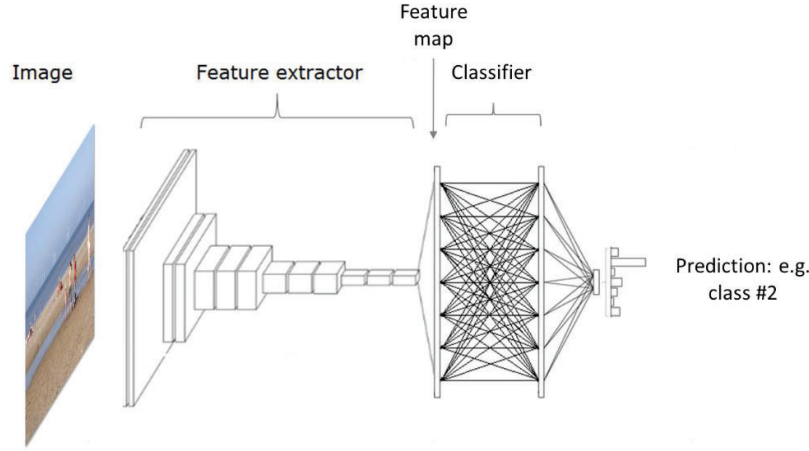141 Finally, an MLP classifier is applied to the feature map in the last layer to produce the final classification
142 result.
143

145 Figure 1: CNN-based architecture
146
147 **2.2. Computer vision-enriched discrete choice models**
148 Throughout this paper, we consider the following choice situation. A decision-maker, n, faces a multi-
149 attribute choice task with a set of J mutually exclusive alternatives. Each alternative is described by M
150 numeric attributes $X_j = \{x_{j1}, x_{j2}, \ldots, x_{jM}\}$, such as e.g. travel cost and travel time and by a (colour) image
151 S_j with a resolution of $H \times W \times C$.
152
153 We assume decision-makers make decisions based on Random Utility Maximising (RUM) principles
154 (McFadden 1974), see Equation 1, where U_{jn} denotes the total utility experienced by decision-maker n
155 considering alternative j, V_{jn} is the utility experienced by decision-maker n derived from attributes
156 observable by the analyst. And, to account for the fact that the analyst does not observe everything that
157 matters to the decision-maker's utility, an additive error term ε_{jn} is added to each alternative (Train
158 2003).
159

$$U_{jn} = V_{jn} + \varepsilon_{jn}$$ Equation 1

160
161 Furthermore, we assume decision-makers experience utility from both the numeric attributes X_j and
162 image S_j, see Equation 2, where v is a preference function which maps the attributes and image of an
163 alternative onto utility.
164

$$U_{jn}\left(X_{jn}, S_{jn}\right) = v\left(X_{jn}, S_{jn}\right) + \varepsilon_{jn}$$ Equation 2

165
166 In addition, we make three more assumptions to develop the CV-DCM:
167
168 1. We assume that the utility derived from the numeric attributes and the image are separable and
169   additive in utility space, see Equation 3, where function f maps the (observed) numeric
170   attributes onto utility and function g maps the relevant information from the image onto utility.
171

$$U_{jn}\left(X_{jn}, S_{jn}\right) = f\left(X_{jn}\right) + g\left(S_{jn}\right) + \varepsilon_{jn}$$ Equation 3

172

4

173   2. We assume that utility is linear and additive with numeric attributes as well as with images'
174   feature maps. Thus, f and g are standard linear-additive utility functions. As discussed in section
175   2.1, feature maps are more compact representations of images. Accordingly, we let $Z_j = \{z_{j1}, z_{j2},$
176       $..., z_{jK}\}$ denote the feature map of image S_j, and $\varphi(w): \mathbb{R}^{H \times W \times C} \to \mathbb{R}^K$ be a function that maps
177   image S_j onto feature map Z_j. Hence, φ is the transformation produced by the feature extractor
178   of a CV model, and w are its associated weights (i.e., parameters). Both the numeric attributes
179   X_j and feature map Z_j enter the utility function in a linear-additive fashion, as shown in Equation
180   4. In Equation 4, β_m denotes the marginal utility associated with attribute m; x_{jmn} denotes the
181   attribute level of numeric attribute m of alternative j, as faced by decision-maker n; and β_k
182   denotes the weight associated with the kth element of feature map Z_{jn}.
183

$$U_{jn} = \underbrace{\sum_m \beta_m x_{jmn}}_{\substack{\text{Utility derived} \\ \text{from numeric attributes}}} + \underbrace{\sum_k \beta_k z_{jkn}}_{\substack{\text{Utility derived} \\ \text{from image feature map}}} + \varepsilon_{jn} \qquad where\ Z_{jn} = \varphi\left(S_{jn} \mid w\right) \qquad \text{Equation 4}$$

184

185   The reason that we let feature maps, as opposed to individual pixel values, enter the utility
186   function is twofold. First, the pixel values in and of themselves are behaviourally meaningless.
187   Rather, the joint distribution of many pixels together produces higher-level behaviourally
188   meaningful concepts that can be expected to generate utility. A feature map can be trained to
189   provide a representation that can be mapped linearly (Goodfellow et al. 2016 pg. 518). Second,
190   a medium-sized colour image of 640 x 480 pixels contains almost 1 million data points. This
191   high dimensionality renders using each pixel directly in the utility function practically
192   infeasible. Instead, $\varphi(w)$ is learned to produce a feature map containing the relevant information
193   of the image to the choice behaviour in a way that maps linearly onto utility.
194

195   3. In line with common practice in choice modelling, we assume ε_{jn} is independent and identically
196   Extreme Value Type I distributed with a variance of $\pi^2/6$, resulting in the well-known and
197   convenient closed-form logit formula for the choice probabilities (P_{in}), given in Equation 5,
198   where C_n denotes the set of alternatives presented to decision maker n. Note that this assumption
199   would, from a machine learning perspective, be equivalent to saying that the output layer is a
200   Softmax function.
201

$$P_{in} = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \qquad\qquad \text{Equation 5}$$

202
203 Figure 2 depicts a graphical representation of the model structure of the proposed CV-DCM model. An
204 essential aspect of the CV-DCM model's architecture is that the network's upper (associated with the
205 left-hand side alternative) and lower (associated with the right-hand side alternative) parts are identical.
206 Thereby, the left and right-hand side alternatives are mathematically treated in the same way, allowing
207 us to interpret the values on the nodes of the last layer as a utility. In the computer vision field, this
208 network structure (bar the numeric attributes) is called a Siamese network (Bromley et al. 1994).
209 Siamese networks are typically used to determine the (dis)similarity between two images. Importantly,
210 even though we can interpret the last layer as utility, we cannot interpret β_k in the same way we can
211 with β_m. β_k can be conceived as a marginal utility –after all, it reflects the change in utility by a unit

212 change in the attribute level. But, because the meaning and units of the elements on the feature map, Z_j,
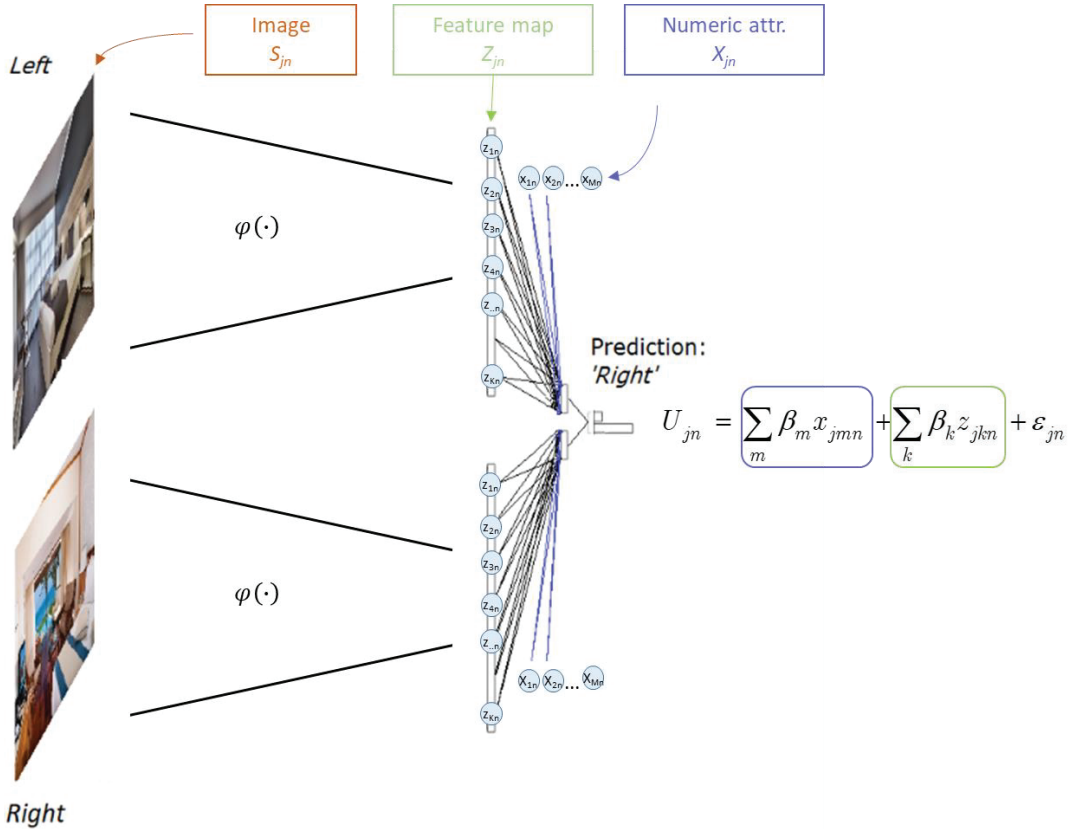213 are unclear, they do not carry a behavioural meaning.
214

215
216 Figure 2: Model structure of CV-DCM
217

2.3. Feature extractor and training

219 In this study, we use the feature extractor of the DeiT base model (Touvron et al. 2021). DeiT models
220 are data-efficient vision transformer-based models which produce competitive capabilities on
221 benchmark data sets, such as ImageNet (Russakovsky et al. 2015), at a lower computational cost and
222 data requirements than many of its competitors (Touvron et al. 2021). The DeiT base model consumes
223 a relatively modest 86 million weights. Furthermore, we use transfer learning to train our CV-DCM
224 (Bengio 2012) to lower the computational time and amount of training data. The idea of transfer learning
225 is to use a pre-trained network as the starting point for developing another network for a closely related
226 task. In other words, rather than retraining the whole model from scratch, we start the training from an
227 already good starting point when we train the CV-DCM. Our pre-trained DeiT base model is trained on
228 ImageNet (Deng et al. 2009); a widely used benchmark image data set containing 1.2 million training
229 images with 1,000 object classes.
230

3. Data collection and training

232 We demonstrate the proposed CV-DCM by applying it to data obtained through a stated choice
233 experiment involving residential location choices. The residential location choice makes a suitable case
234 study because both numeric attributes and visual information can be expected to be important to
235 residential location choice behaviour (Smith and Olaru 2013). Moreover, images of the sort that we
236 need for modelling residential location choice behaviour, namely street-view 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-view 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. Figure 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 (Perez-Lopez et al. 2022). Secondly, they apply generically to almost everyones' 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.

Figure 3: Screenshot of the pivoted stated choice experiment
(translated to English; original in Dutch)

As can be seen in Figure 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.

Table 1: Attribute levels Stated Choice experiment

Current commute travel time of the respondent (TT_n)	Attribute levels	
	Housing cost (hhc) [€]	Commute travel time (tti) [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

3.1.1. Street-view 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-view images we created before conducting the stated choice experiment. A major effort went into the construction of this database with street-view 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

294   images taken in 2020, or later. Each image id corresponds to a 360-degree panorama photo. Fourth,
295   from each panorama, we generated two image urls with 90-degree angles to the direction of the street
296   (to both directions). This latter ensures the images are 'window views' (e.g. as opposed to views parallel
297   to the driving direction of the Google car). Finally, urls of images of poor quality were algorithmically
298   removed. More specifically, urls to black images, blurred images and images with tilted horizons were
299   removed. The final database contains the urls a little over 60k street-view images of residential streets
300   from 50 municipalities in the Netherlands.

301

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

308

3.1.2. Experimental design

309

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

323

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

333

3.2. Data collection and sample description

334

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

342 commute days). Noteworthy, we also asked respondents how important the three attributes (housing
343 cost, travel time and the image) were for their decisions on a scale from 1 to 10. Although it is well-
344 known that direct elicitation of preferences is treacherous (Nisbett and Wilson 1977), it still can provide
345 first (albeit inconclusive) evidence of the importance of the presented street-view images relative to the
346 numeric attributes for the residential location choices.
347
348 The target population for the survey was the Dutch population of 18 years and older, with ten or more
349 minutes of commute travel time. The latter requirement was necessary because we used a pivoted
350 experimental design. Because of this latter condition, no official population statistics exist to compare
351 our sample against, but we do not expect this condition to affect the population statistics substantially.
352 Therefore, care was taken that the sample was, by and large, representative of the Dutch 18 years and
353 older population in terms of gender, age and spatial distribution across the Netherlands. Cint[1], a panel
354 data provider, provided the panel of respondents. In total, 800 respondents completed our survey.
355
356 Table 2 shows the sample statistics. Overall, the sample is representative of the target population. Also,
357 for the variables that are not explicitly considered during the data collection, such as e.g. the modal split
358 and household composition, the statistics are close to the population data (c.f. Ton et al. 2019).
359 Furthermore, looking at the reported monthly housing cost, we notice that the largest share of the
360 respondents has a housing cost below €750. This seems reasonable since the average net housing cost
361 of rental houses in the Netherlands is around €700 p/m; homeowners' average net housing cost is slightly
362 above €900 p/m (Stuart-Fox et al. 2022).
363
364

Table 2: Sample statistics

Socio-demographic variable	Category	Distribution
Age	18 - 29 year	21%
	30 - 39 year	19%
	40 - 49 year	20%
	50 - 59 year	22%
	60 - 69 year	17%
	+70 year	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%

---

[1] See www.cint.com

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%

3.3. Descriptive analysis

Figure 4 shows histograms of the self-reported importance levels of the street-view images (left), monthly housing costs (middle) and commute travel times (right). Figure 4 shows that the street-view images 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-view images 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 most variance for the commute travel time. Noteworthily, the importance rating for the street-view images is weakly negatively correlated with the importance ratings for monthly housing costs ($\rho = -0.10$) and uncorrelated with the ratings for commute travel time ($\rho = 0.02$). In contrast, the importance ratings for monthly housing costs and commute travel times are strongly positively correlated ($\rho = 0.36$). This strong positive correlation reveals that people who find housing costs important usually also find commute travel time important, and vice versa.

Figure 4: Self-reported importance levels of attributes in the SC experiment

Figure 5 shows the Pearson correlation coefficients between importance ratings and a selection of respondent characteristics. Interestingly, the top row shows that the importance of the street-view images 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-view images 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., Bhat and Guo 2007; 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 the street-view images.

Furthermore, Figure 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 Figure 5 seem plausible.
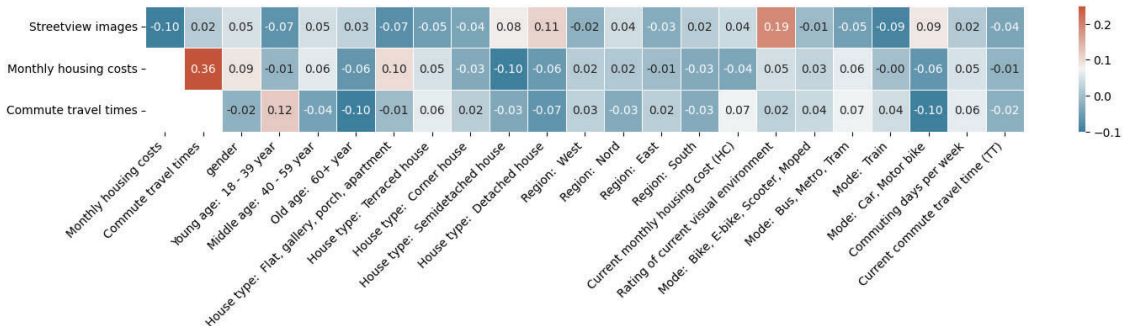

Figure 5: Pearson correlation coefficients between importance ratings and respondent characteristics

12

408 Next, we analyse the images used in the stated choice experiment. Although our street-view image
409 database comprises urls to over 60k images, only slightly over 7.5k unique images are used in the stated
410 choice experiment. Because images are drawn randomly from our image database with replacement,
411 we expect that some images will be sampled more than once. Indeed, most images are used once.
412 However, counter to our design intentions, some images are used 20 times or more. A possible
413 underlying cause could be the seed numbers used by the survey platform's software. Nevertheless,
414 regardless of this issue's origin, when we deal with the issue carefully during the training of our models
415 (see Section 3.4.4), it will not have to impact our (substantive) findings.
416
417 Finally, Figure 6 shows the distribution of the month of the year of the images used in the survey. In
418 line with expectations, the images are not evenly distributed over the year. We see that most images are
419 taken in spring and summer (March to September). Furthermore, we notice that images are sampled for
420 all 12 months. This implies we can account for the impact of the seasons on the utility derived from the
421 street-view images by estimating constants for all months (except one, which we need to fix to zero for
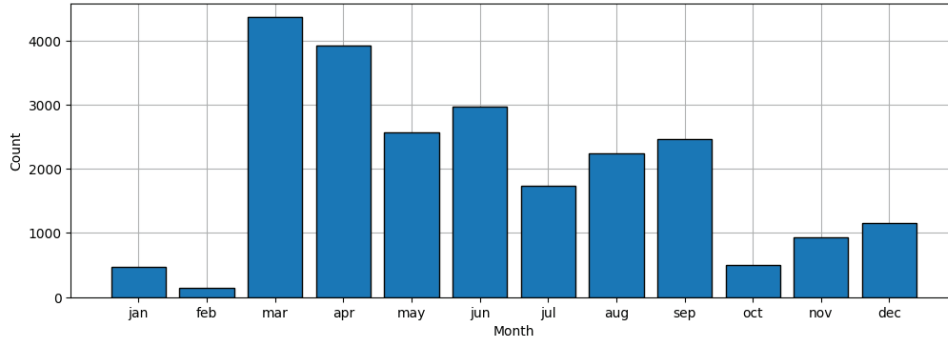422 normalisation).
423

424
425   Figure 6: Distribution of images used in the stated choice experiment over the months of the year
426

3.4.   Training
3.4.1. Loss function and implementation
429 Training a CV-DCM involves finding the weights of the model (β, w) that minimise the loss function.
430 In other words, the weights of the feature extractor and preference parameters of the utility function are
431 jointly optimised. For this study, we use a cross-entropy loss function with an L2 regularisation term,
432 see Equation 6. Minimising the cross-entropy loss is equivalent to maximising the Log-Likelihood (LL)
433 of the data given the model – which is common practice in the choice modelling literature. The L2
434 regularisation aims to reduce the chance of model overfitting by penalising the magnitude of the weights
435 in the model. γ governs the strength of the regularisation. Note that we apply regularisation only to w
436 and not to preference parameters β_m and β_k. Regularising preference parameters could lead to
437 undesirable biases.
438

$$w^*, \beta^* = \arg\min_{w,\beta} \left[\overbrace{\frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{J} y_{nj} \log\left(P_{nj} \mid X_{nj}, S_{jn}, \beta \right)}^{\text{cross-entropy loss}} + \overbrace{\gamma \sum_{r=1}^{} w_r^2}^{\text{L2 regularisation}} \right] \qquad \text{Equation 6}$$

439

440 The model's source code and data are openly available.[2] Thereby, we aim to support model-building
441 and validation practices. We hope our data can become a benchmark data set for studying choice
442 behaviour in the presence of visual stimuli.
443
444 **3.4.2. Implementation and hyperparameter tuning**
445 Our CV-DCM is implemented and trained in PyTorch (Paszke et al. 2019). PyTorch is a Python-based
446 machine learning package commonly used for deep learning computer vision research because it
447 supports GPU computing. We have determined the hyperparameters of the CV-DCM using a 'heuristic
448 search approach'. We have tried various combinations of optimisation algorithms, learning rates, batch
449 sizes, and regularisation settings to determine the 'optimal' hyperparameters. Ideally, we would have
450 conducted a full-fledged hyperparameter tuning in which all combinations of optimisation algorithms,
451 learning rates, batch sizes, and regularisation parameters would be tested. But, given the computational
452 cost of training CV-DCM (and CV models more generally), we refrained from conducting such a
453 hyperparameter tuning. We found the following hyperparameters to work best (Table 3).
454
455

Table 3: Hyperparmeters CV-DCM

Hyperparameter	Value
Optimisation algorithm	Stochastic Gradient Descent
Batch size	20
L2 weight decay (γ)	0.1
Learning rate	1e-6

456
457 **3.4.3. Image transformation and feature scaling**
458 In line with common practice in computer vision, we transform and augment images while training the
459 CV-DCM. Specifically, we conduct the two operations. First, we downsampled the images to 224 x 224
460 pixels. This downsizing operation ensures that images have the input dimensions expected by the CV
461 model (i.e. DeiT base model). Second, we randomly flip images horizontally. This data augmentation
462 operation reduces the model's ability to remember images, thus lowering the chance of overfitting the
463 training data. Furthermore, we scale the numeric features. Scaling the features helps the optimiser to
464 avoid getting stuck in local minima (Géron 2019). The most common type of scaling in machine
465 learning involves shifting and scaling the features to a zero mean and a unit variance. We use another
466 commonly used scaling technique to scale the housing cost and travel time features, called min-max
467 scaling. This scaling entails scaling the features to a range of [-1,1]. The advantage of this scaling
468 technique is that it is straightforward and facilitates easy interpretation of the model's parameters. To
469 facilitate interpretation, we have used the same scaling for all data (thus ignoring that the minimum
470 travel time level varied across respondents, see Table 1). Specifically, all housing costs are divided by
471 225 and travel times are divided by 15.
472
473 **3.4.4. Train-test split**
474 Splitting the data into a train set and a test set is essential for training virtually all machine learning
475 models because their high capacity makes them prone to overfitting (Géron 2019). As the name
476 suggests, the train set is used for training the model; the test is unseen by the model during training and
477 used to evaluate (test) the model's generalisation performance after training. If a trained model overfits
478 the data, a gap in the performance between the train and test set will tell.
479

---

[2] github.com/sandervancranenburgh/Computer-vision-enriched-DCMs

480 The most common way to create the train-test split is by randomly allocating observations to the two
481 sets. When splitting data, it is important to avoid "data leakage". Data leakage happens when the model
482 has access to information during training that it does not have when deployed after training (see e.g.
483 Hillel 2021 for its impact on choice model outcomes). For this study, we split our data across images.
484 Thereby, we aim to avoid potential data leakage from learning the utility levels of specific images rather
485 than generalisable high-level utility-generating features embedded in the images. Making such a split
486 is, however, a nontrivial network problem. Every image is connected at least to one other image (the
487 other street-view image presented in the choice task). However, some images are connected to dozens
488 of other images because they are used more than once (see section 3.3). Hence, when we assign one
489 image to the train set, we must place all directly and *indirectly* connected images in the train data set
490 too.
491
492 Given the above 'network' problem, we took the following procedure to create the train and test sets.
493 We randomly picked one choice task, comprising two images, and put this choice task and *all* choice
494 tasks connected to this one in the train set. We repeated the random picking of choice tasks until 80%
495 of the data were used. The remaining data (20%) make the test data set. The train and test data sets
496 comprise, respectively, $N = 9,784$ and $N = 1,948$ choice observations. Due to our splitting strategy,
497 observations of the same individual may be present in both the train and test data sets. However, it is
498 unlikely to cause serious data leakage because no socio-demographic variables (that would be needed
499 to identify observations of the same respondent) are used in the training of the CV-DCM.
500

4. Results

4.1. Estimation results

503 Table 4 shows the main estimation results for the three models, whose utility functions are given in
504 Equation 7 to Equation 9. Models 1 and 2 are standard linear-additive RUM-MNL models used as
505 benchmark models to compare the proposed CV-DCM (Model 3). All models assume decision-makers
506 experience (dis)utility from the housing cost and travel time in a linear and additive fashion. Model 1
507 ignores the images completely, while Model 2 takes into account the month in which the image is taken
508 by estimating constants, denoted β_{mo}, for each month. If where and when images are collected are
509 uncorrelated, we expect that images taken in spring and summer, on average, attain a higher utility than
510 images taken in autumn or winter. Model 3 is the proposed CV-DCM and takes the monthly housing
511 cost (hhc), commute travel time (tti) and the month of the year as numeric input attributes in the same
512 way as Model 2 does, but also takes the feature maps of the images as inputs. For each model, Table 4
513 shows the performance on the train and test sets, using three (related) metrics: the log-likelihood, rho-
514 square and cross-entropy. A good performance on the test set implies the model generalises well to
515 unseen data. A gap between the performance on the train and test set signals overfitting.
516

$$U_{in} = \beta_{hhc}hhc_{in} + \beta_{tti}tti_{in} + \varepsilon_{in} \qquad\qquad \text{Model 1} \qquad\qquad \text{Equation 7}$$

$$U_{in} = \beta_{hhc}hhc_{in} + \beta_{tti}tti_{in} + \sum_{mo}\beta_{mo}I_{S_j} + \varepsilon_{in} \qquad\qquad \text{Model 2} \qquad\qquad \text{Equation 8}$$

$$U_{in} = \beta_{hhc}hhc_{in} + \beta_{tti}tti_{in} + \sum_{mo}\beta_{mo}I_{S_j} + \sum_{k}\beta_{k}z_{ikn} + \varepsilon_{in} \qquad \text{Model 3} \qquad\qquad \text{Equation 9}$$

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

$$z_{ikn} = \phi\left(S_{in} \mid w_r\right),$$

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

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-view images to predict the choice behaviour. Looking at the generalisation performance, we see that CV-DCM outperforms the two benchmark models by a large margin. Specifically, the CV-DCM improves the log-likelihood on the test set by 57 log-likelihood points. The rho-square jumps from 0.116 to 0158. 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 equally well on the test set. Hence, the plain incorporation of the month of the year in the utility function does not improve the generalisability of the standard RUM-MNL models. Third, despite having to train 86 million weights, the extra computational time does not render the CV-DCM impractical; 4.5 hours of training time is in the same order of magnitude as the estimation time of conventional mixed logit models. But, handling large numbers of images and working with GPUs is technically considerably more challenging.

Next, we look at the estimated preference 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.[3] In the context of our SC experiment, the VTT gives the (mean) willingness to pay per month for a one-hour travel time reduction. 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 the three models. We expect stable β_{tti}/β_{hhc} ratios because our experimental design is constructed such 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 with images taken in December (which we fixed to zero). In Model 2, we see that the estimates associated with the months of the year are mostly positive and significant for the spring and summer months. This can be explained by the notion that images taken in these months are more likely to look more attractive to live than images taken in winter, for instance, because the weather is better. However, the positive and significant

---

[3] $VTT = 60\left(\dfrac{225}{15}\right)\dfrac{\beta_{tti}}{\beta_{hhc}}$ Note that the factor (225/15) comes from the fact that the attributes are scaled before training.

551 estimate for January counters this line of argumentation and is hard to explain. Now we turn to the
552 estimates associated with the months of the year under Model 3. In Model 3, the estimates associated
553 with the months of the year do not carry the same interpretation as under Model 2. The utility derived
554 from an image in Model 3 is the sum of the utility from the image's feature map and the estimate for the
555 month of the year. As a result, we cannot see the estimates associated with these two utility sources in
556 isolation. One noteworthy observation concerning the estimates related to the months of the year in
557 Model 3 is that fewer estimates are significant than in Model 2. This observation aligns with statistical
558 expectations. Because feature maps already contain information about the month of the year, adding the
559 month of the year explicitly to the model provides little additional information to explain the choice
560 behaviour. To give an example, an image in which trees that have shed their leaves reveals the image
561 is probably taken in winter.

562

563 Table 4: Estimation results

		Model 1			**Model 2**			**Model 3**		
Model type		lin-add RUM-MNL			lin-add RUM-MNL			CV-DCM		
Number of parameters		2			13			86m		
Estimation time		<1 sec[I]			<1 sec[I]			4.5 hr.[II]		
Train set N = 9784	Log-Likelihood	-5,954			-5,931			-5,724		
	ρ^2	0.120			0.130			0.156		
	Cross-entropy	0.609			0.606			0.585		
Test set N = 1948	Log-Likelihood	-1,194			-1,194			-1,137		
	ρ^2	0.116			0.116			0.158		
	Cross-entropy	0.613			0.613			0.585		
		est	s.e.	p-val	est	s.e.	p-val	est	s.e.[III]	p-val[III]
β_{hhc}		-0.86	0.025	0.00	-0.87	0.024	0.00	-0.96	0.025	0.00
β_{tti}		-0.21	0.023	0.00	-0.21	0.025	0.00	-0.24	0.026	0.00
β_{jan}					0.46	0.129	0.00	0.25	0.136	0.07
β_{feb}					0.02	0.228	0.91	-0.40	0.240	0.10
β_{mar}					0.10	0.080	0.23	0.05	0.084	0.58
β_{apr}					0.25	0.080	0.00	0.36	0.084	0.00
β_{may}					0.28	0.084	0.00	0.08	0.088	0.39
β_{jun}					0.17	0.084	0.04	-0.12	0.088	0.16
β_{jul}					0.21	0.094	0.02	0.31	0.098	0.00
β_{aug}					0.24	0.087	0.01	0.12	0.092	0.17
β_{sep}					0.19	0.085	0.03	0.33	0.089	0.00
β_{oct}					0.46	0.131	0.00	0.40	0.138	0.00
β_{nov}					-0.11	0.106	0.31	-0.04	0.111	0.74
β_{dec}					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

[I] Using 4 CPUs (Xeon @ 3.60 GHz)

[II] Using GPU (GeForce RTX 2080Ti)

564 [III] Obtained though computing the hessian while keeping the utility derived from the image fixed

565

566 Lastly, we analyse the contributions to utility differences between the right and left-hand side
567 alternatives derived from the images' feature maps. To do so, Figure 7 shows three kernel density plots.
568 The left-hand side plot shows the total utility difference as predicted by the trained CV-DCM; the
569 middle plot shows the utility difference from the numeric attributes; and the right-hand side plot shows
570 the utility difference from the images. We make several observations based on Figure 7. Firstly, looking
571 at the range of x-axes of the middle and right-hand side plots, we see that the utility differences arising
572 from the images and numeric attributes are similar. This tells us that the part-worth utilities derived
573 from the numeric attributes (housing cost and travel time) are of the same magnitude as those derived

574 from the street-level conditions embedded in the street-view images.[4] This observation adds to the
575 evidence that street-level conditions are important to residential location choice behaviour and can
576 effectively be captured and modelled using images and CV-DCMs. Secondly, we notice that the
577 distributions of utility differences are virtually equal for the test and train sets. This indicates that the
578 CV-DCM does not overfit the training data, and the data are adequately split into train and test sets.
579 Therefore, the CV-DCM must have learned to extract salient generalisable features from the images
580 that generate utility. Thirdly, we see that the distribution of the utility differences stemming from the
581 images is comparatively more bell-shaped than those of the numeric attributes. At first sight, this may
582 seem odd, but it can be explained by how the choice tasks are constructed. Recall that we removed
583 choice tasks without trade-offs between the numeric attributes (see section 3.1.2 for more details). This
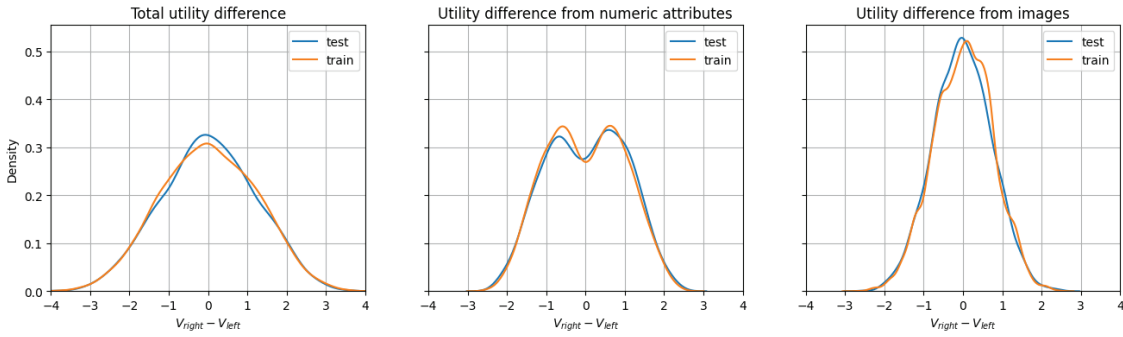584 removal leads to the bi-modal shape of the utility difference.
585

Figure 7: Utility differences

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

590 The considerable performance improvement by the CV-DCM compared to the benchmark models (see
591 Table 4) supports the notion that the CV-DCM can extract relevant information from images to predict
592 choice behaviour. But, β_k and w do not carry a behavioural meaning. Therefore, they do not provide
593 directly interpretable insights about what the CV-DCM has learned regarding what decision-makers
594 find important for their residential location choices. To shed light on what the CV-DCM has learned
595 about decision-makers' preferences, we show two collages of images taken from the test set, to which
596 the trained CV-DCM assigns the highest (Figure 8) and lowest (Figure 9) utility levels. Note that the
597 utility level is stamped in the top left of each image.[5] These utility levels are 'uncorrected' for the month
598 of the year. Hence, the top left image yields a utility of 1.63 if the image was taken in December, while
599 it produces a utility of 1.63-0.12 = 1.51 if it was taken in August (which it is).
600
601 What catches the eye in Figure 8 is that the images all look spacious, green and often water-abundant.
602 We see many trees, grassland and detached houses. In the authors' view, these street-view images indeed
603 are highly attractive residential locations. In sharp contrast, the images in Figure 9 look cramped,
604 greyish, and urbanised and often have hallmarks of transportation, such as overhead wires, bus stops,
605 parked bikes, and cars. In the authors' view, these street-view images are highly unattractive as
606 residential locations. The mean difference in utility between the 20 best, shown in Figure 8, and the 20
607 worst images, shown in Figure 9, is 2.7 utility points. The willingness to pay per month to move from
608 the worst to the best street-level conditions can be computed by dividing the utility difference by β_{hhc}.

---

[4] Given the ranges of the numeric attributes presented in the SC experiment.
[5] Note that the mean utility derived from images across all images is 0.02 (thus not exactly zero). This is inconsequential as utility has no absolute scale of level, only utility differences matter (Train 2003).

609 The result yields a willingness to pay of 632 euros per month – which seems plausible in the authors'
610 view.
611

612
613 Figure 8: Images with the highest predicted utility levels
614
615

616
617 Figure 9: Images with the lowest predicted utility levels
618

619 **4.3. Heterogeneity in residential location choice preferences**
620 To further demonstrate the merits of CV-DCMs for modelling residential location choice behaviour
621 and, more generally, modelling choice behaviour in the context of visual and numeric stimuli, we show
622 CV-DCMs can be used to investigate preference heterogeneity. For this analysis, we use the Latent
623 Class (LC) modelling framework (Greene and Hensher 2003; Hess 2014). Because the class
624 membership function in LC models can be parameterised based on the socio-demographic
625 characteristics of decision-makers, they are particularly well-suited to uncover profiles of groups with

626 different preferences. Equation 10 shows the general log-likelihood function of the (panel) LC model,

627 where $P\left(i_{nt} \mid \beta_s\right)$ is the probability predicted by class s of the chosen alternative in choice task t of

628 decision-maker n. In LC models, a class membership function, $f\left(g_n \mid \gamma\right)$, probabilistically allocates

629 respondents to different classes, and each class s has its own set of preference parameters

630 $\beta_s = \left\{\beta_{1s}, \beta_{2s}, ..., \beta_{Ms}\right\}$. The probability decision-maker n belongs to class s, denoted π_{ns}, is thus a

631 function of the individual's characteristics, $g_n = \left\{g_{n1}, g_{n2}, ..., g_{nQ}\right\}$, and estimable parameters

632 $\gamma = \left\{\gamma_1, \gamma_2, ..., \gamma_Q\right\}$.

633

$$LL\left(\beta, \gamma \mid X, G\right) = \sum_{n=1}^{N} \ln\left[\sum_{s=1}^{S} \pi_{ns} \prod_{t=1}^{T_n} P\left(i_{nt} \mid \beta_s\right)\right]$$
$$where\ \pi_{ns} = f\left(g_n \mid \gamma\right)$$

Equation 10

634

635 The estimation of a CV-DCM in LC form is technically beyond reach. The already high number of

636 weights (86m) would further increase, scaling linearly with the number of classes. To avoid the number

637 of weights exceeding technical constraints[6], we freeze w and β_k to their trained values. Consequently,

638 the average utility level of image S_{int}, denoted v_{int}, can be computed before estimation and enters the

639 utility function directly as an explanatory attribute, see Equation 11. The associated preference

640 parameter, β_{emb}^s, represents the importance given to the images' utility levels within class s.

641

$$P\left(i_{nt} \mid \beta_s\right) = \frac{e^{V_{int}}}{\sum_j e^{V_{jnt}}}, \quad where \quad V_{int} = \beta_{hhc}^s hhc_{int} + \beta_{tti}^s tti_{int} + \beta_{emb}^s v_{int},$$
$$v_{int} = \sum_k \beta_k \varphi\left(S_{int} \mid w\right)$$

Equation 11

642

643 We have estimated CV-DCMs in LC form using Apollo (Hess and Palma 2019) for two to five latent

644 classes. In line with the practices in machine learning, the optimal number of classes is determined

645 based on generalisation performance (as opposed to statistical metrics, such as BIC and AIC) (Parady

646 et al. 2020). To do so, we applied the trained LC models to our test data set and determined the log-

647 likelihood of the test set (i.e. LL_{test}). Although the 4-class model marginally outperforms the 3-class LC

648 model, we choose to use the 3-class model for further development for reasons of interpretability.

649 Specifically, we parameterised the class membership function of the 3-class model using the

650 respondents' characteristics, g_n, reported in Table 2. To determine the final set of respondent

651 characteristics, for each characteristic we tested whether it significantly contributed to the membership

652 function, at least in one class, using a critical level of significance of $\alpha = .10$.

653

654 Table 5 shows the estimation results of the 3-class LC model. We make the following inferences based

655 on Table 5. Firstly, accounting for preferences heterogeneity increases the model performance

656 substantially. Specifically, the log-likelihood on the test data sets drops 77 LL points compared to Model

657 3 – the best single-class model (see Table 4). Accordingly, the rho-square jumps from 0.158 to 0.215.

---

[6] e.g. available RAM on the GPU

Secondly, the three latent classes represent three highly distinct sets of preferences. Specifically, respondents typical to Class A (the largest class) consider the street-level conditions and commute travel time to be the most important attributes and monthly housing costs to be the least important. Respondents typical to Class B consider monthly housing cost to be of greatest importance, attributing comparatively little importance to street-level conditions and commute travel times. Respondents typical to Class C seem to be only sensitive to the monthly housing costs. But, the small sizes of all the estimates within this class indicate respondents in this class display highly erratic choice behaviour.

Table 5: LC estimation results

	Model 4										
Model type	CV-DCM LC model										
Number of parameters	35										
Estimation time	1.5 min[I]										
Train set N = 9784 — Log-Likelihood	-5,158										
Train set N = 9784 — ρ^2	0.239										
Train set N = 9784 — Cross-entropy	0.527										
Test set N = 1948 — Log-Likelihood	-1,060										
Test set N = 1948 — ρ^2	0.215										
Test set N = 1948 — Cross-entropy	0.544										
	Class A [45%] Residents with high WTP for street level factors and high VTT				**Class B [41%]** Residents with low WTP for street level factors and moderate VTT				**Class C [14%]** Residents only caring about monthly cost			
Class description											
Choice model estimates	*est*	*s.e.*	*t-val*	*p-val*	*est*	*s.e.*	*t-val*	*p-val*	*est*	*s.e.*	*t-val*	*p-val*
β_{hhc}	-0.51	0.056	-9.07	0.00	-2.59	0.138	-18.84	0.00	-0.20	0.091	-2.19	0.03
β_{tti}	-0.24	0.050	-4.75	0.00	-0.54	0.068	-7.86	0.00	-0.14	0.094	-1.47	0.14
β_{emb}	1.58	0.094	16.79	0.00	0.86	0.084	10.28	0.00	0.08	0.142	0.59	0.56
VTT[II]	€422	101.9		0.00	€186	31.4		0.00	€628	637.1		0.32
WTP$_{street\text{-}level\ factors}$[II]	€701	114.2		0.00	€75	10.8		0.00	€95	187.8		0.61
Class membership model estimates												
δ	0.00	--fixed			0.36	0.557	0.64	0.52	0.57	0.980	0.58	0.56
γ_female	0.00	--fixed			0.34	0.220	1.53	0.13	-0.77	0.414	-1.86	0.06
γ_age_young (age≤39yr.)	0.00	--fixed			0.00	--fixed			0.00	--fixed		
γ_age_middle (40yr.≤age≤59yr.)	0.00	--fixed			-0.37	0.245	-1.51	0.13	-1.73	0.511	-3.39	0.00
γ_age_old (60yr.≤age)	0.00	--fixed			-0.28	0.313	-0.90	0.37	-2.37	0.799	-2.97	0.00
γ_housetype_flat_porch_apartment	0.00	--fixed			0.00	--fixed			0.00	--fixed		
γ_housetype_terraced_house	0.00	--fixed			-1.03	0.339	-3.03	0.00	-1.18	0.575	-2.05	0.04
γ_housetype_corner_house	0.00	--fixed			-1.07	0.385	-2.78	0.01	-0.40	0.603	-0.67	0.51
γ_housetype_semidetached_house	0.00	--fixed			-1.75	0.407	-4.30	0.00	-1.58	0.766	-2.07	0.04
γ_housetype_detached_house	0.00	--fixed			-2.03	0.412	-4.93	0.00	-2.52	0.790	-3.19	0.00
γ_region_West	0.00	--fixed			0.00	--fixed			0.00	--fixed		
γ_region_Nord	0.00	--fixed			0.10	0.387	0.27	0.79	0.30	0.580	0.51	0.61
γ_region_East	0.00	--fixed			0.13	0.272	0.50	0.62	-1.87	0.745	-2.51	0.01
γ_region_South	0.00	--fixed			-0.06	0.279	-0.22	0.83	-0.64	0.555	-1.15	0.25
γ_commuting_days_per_week	0.00	--fixed			0.15	0.081	1.83	0.07	0.20	0.158	1.25	0.21

[I] Using 4 CPUs (Xeon @ 3.60 GHz)

[II] $VTT = 60\left(\frac{225}{15}\right)\frac{\beta_{tti}}{\beta_{hhc}}, WTP = -225\frac{\beta_{emb}}{\beta_{hhc}}$

Now we turn to the class membership model. What catches the eye is that several characteristics that correlate strongly with the self-reported importance ratings (shown in Figure 5) are not in Table 5. For instance, based on the correlations shown in Figure 5, we would expect the rating of the current visual street-level conditions and current mode of transport to co-explain the class membership of decision-makers. But, these characteristics do not significantly affect the class membership and thus are dropped from the final model reported in Table 5. Instead, Table 5 shows that age and house type are strongly associated with class membership.

To further facilitate the interpretation of the membership model, Figure 10 visualises the across-class (left) and within-class profiles (right). As the name suggests, the across-class profile shows the distribution for each characteristic level across the classes. In the across-class profile, the probabilities in each row sum up to one. The across-class profile, for instance, tells us how males are distributed across the three classes. The within-class profile shows the distribution across respondent characteristics (e.g. gender) within a class. It sheds light on the share of, e.g. males and females across the respondents

21

within a given class. Hence, in the within-class profile, the probabilities column-wise sum up to one (for each characteristic).

Based on Figure 10, we make several observations. Firstly, the across-class profile plot shows that most males are in Class A, while most females are in Class B. But, the within-class profile also reveals that males are particularly overrepresented in Class C. Secondly, most middle and older respondents are in Class A, while most young respondents are in Class B. Looking at the within-class profiles, we see that young respondents are particularly overrepresented in Class C. Thirdly, most respondents living in a flat, gallery, porch or apartment are in Class B, while respondents living in semi-detached and detached houses are overrepresented in Class A. Fourthly, and somewhat surprisingly, we see a noteworthy difference between the four regions. Specifically, unlike respondents living elsewhere, most respondents living in the West of the Netherlands are in Class B. Moreover, respondents living in the West of the Netherlands are strongly overrepresented in Class C. Finally, most respondents commuting three days or less are in Class A, while most respondents commuting more than three days are in Class B. Looking at the within-class distribution, we observe that respondents commuting five days are particularly overrepresented in Class C.

Altogether, a coherent picture emerges: respondents in Class A seem to be older, living in more expensive house types and commuting less frequently (possibly because of working-from-home practices). They have a comparatively high willingness to pay for street-level conditions and a high Value-of-Travel-Time. Respondents in Class B mostly live in flats, galleries, and porches and commute more often than respondents in Class A. They have a comparatively lower willingness to pay for street-level conditions and travel time savings than respondents in Class A. Finally, respondents in Class C are primarily young, commute four or five days a week, and live in the West of the Netherlands. They have little interest in street-level conditions.
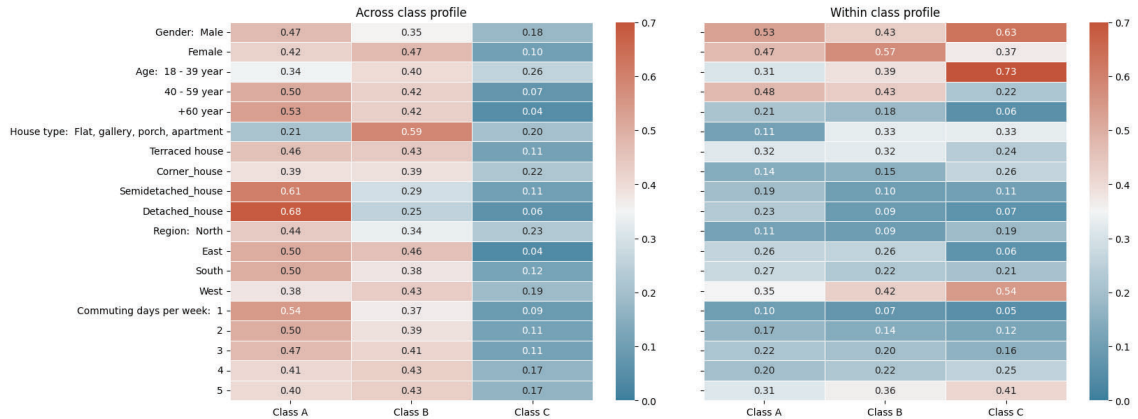
	Across class profile			Within class profile		
	Class A	Class B	Class C	Class A	Class B	Class C
Gender: Male	0.47	0.35	0.18	0.53	0.43	0.63
Female	0.42	0.47	0.10	0.47	0.57	0.37
Age: 18 - 39 year	0.34	0.40	0.26	0.31	0.39	0.73
40 - 59 year	0.50	0.42	0.07	0.48	0.43	0.22
+60 year	0.53	0.42	0.04	0.21	0.18	0.06
House type: Flat, gallery, porch, apartment	0.21	0.59	0.20	0.11	0.33	0.33
Terraced house	0.46	0.43	0.11	0.32	0.32	0.24
Corner_house	0.39	0.39	0.22	0.14	0.15	0.26
Semidetached_house	0.61	0.29	0.11	0.19	0.10	0.11
Detached_house	0.68	0.25	0.06	0.23	0.09	0.07
Region: North	0.44	0.34	0.23	0.11	0.09	0.19
East	0.50	0.46	0.04	0.26	0.26	0.06
South	0.50	0.38	0.12	0.27	0.22	0.21
West	0.38	0.43	0.19	0.35	0.42	0.54
Commuting days per week: 1	0.54	0.37	0.09	0.10	0.07	0.05
2	0.50	0.39	0.11	0.17	0.14	0.12
3	0.47	0.41	0.11	0.22	0.20	0.16
4	0.41	0.43	0.17	0.20	0.22	0.25
5	0.40	0.43	0.17	0.31	0.36	0.41

Figure 10: Across and within class profiles from latent class model

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", for modelling multi-attribute choice behaviour in the presence of visual and numeric stimuli – methodologically expanding the realm of discrete choice models. The proposed Computer Vision-enriched Discrete Choice Model

(abbreviated as 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 and showing it produces new insights into preferences over visual street-level conditions. Notably, we have uncovered which residential places people find most and least attractive and which type of people are particularly willing to pay to improve street-level conditions in their residential area and which are not.

The proposed model, in conjunction with SC experiments, can potentially enhance the travel behaviour research field's understanding of various other transport-related preferences. 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 can 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), our trained CV-DCM could be deployed to assess the spatial distribution of the attractiveness of large residential areas by feeding all available street-view images and visualising how utility is spatially distributed.

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. Three important questions are of particular interest: (1) how to handle multiple images, (2) how to extract more and better information from the trained CV-DCM and (3) how to effectively employ images and ensure that the features of interest are learned from the images (.e.g as oppose the quality of the image, which would instead reflect the photographer's skills).

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. 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. The computer vision field can provide good starting points to address this research question. More specifically, 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. They can shed light on what features are learned by the model to explain the choice behaviour and help validate CV-DCMs. Insights provided by XAI 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 Figure 9) or unattractive (as shown in Figure 10), which can inform the development of policies.

Another area of research that shows promise for extracting additional insights from CV-DCMs is explicitly modelling preference heterogeneity over visual features. Thereby, a more nuanced understanding of the importance of visual features to choice behaviour can be obtained. Because

training the CV-DCM in a fully parameterised latent class form was technically infeasible, we froze the part of the model that mapped the images onto utility in our analysis. Consequently, we could not capture the heterogeneity in preferences over visual aspects. However, different age groups or life stages may appreciate other aspects of the visual environment. For instance, young adults may value vibrancy, while middle-aged people with children may place more importance on street safety. One potential approach to address this challenge is to use a more lightweight feature extractor in the CV-DCM, such as TNT-Ti (Han et al. 2021), DeiT-Tiny (Touvron et al. 2021), and Mixer-S/16 (Tolstikhin et al. 2021). Relatedly, improving and enriching the utility specification of CV-DCMs is a promising avenue for further research. For instance, interaction effects between the numeric and visual information, reference effects, and visual correlates of anchoring effects could be modelled (Hevia-Koch and Ladenburg 2019).

The third question pertains to how to effectively use images in SC experiments such that information relevant to researchers can be extracted. The iconic modern art painting "*Ceci n'est pas une pipe*" neatly illustrates this point. 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). In the context of this study's application, we want to extract information from the image about the attractiveness of the residential location instead of, e.g. which image is most scenic or has the best lighting conditions. Keys to the effective use of images in SC 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. For example, cognitive psychologists have long studied the congruity effect (Henik and Tzelgov 1982). The congruity effect describes the tendency for responses to be faster and more accurate when the items being compared are semantically congruent, such as when the word "red" is depicted in red font. In choice experiments, congruity is occasionally used, albeit not well understood. Specifically, in choice experiments, green and red fonts are frequently used to signal a positive or negative direction of an attribute change, with the idea of lowering the cognitive burden for the respondent. Further investigation is needed to determine whether and under what circumstances preferences obtained from discrete choice models are influenced by the medium of information and how notions from cognitive psychology, like congruency, can be leveraged for better preference elicitation.

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 choice modelling, computer vision and cognitive psychology can accelerate progress. By sharing our data and code openly, we hope to contribute to this advancement.

References

Arkoudi, I., Azevedo, C. L. & Pereira, F. C. (2021). Combining Discrete Choice Models and Neural Networks through Embeddings: Formulation, Interpretability and Performance. *arXiv preprint arXiv:2109.12042*.

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.

Bhat, C. R. & Guo, J. Y. (2007). A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B: Methodological*, *41(5)*, 506-526.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. (1994). Signature verification using a" siamese" time delay neural network. Advances in neural information processing systems.

Brownstone, D., Bunch, D. S. & Train, K. (2000). Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research Part B: Methodological*, *34(5)*, 315-338.

Cao, X. (2014). Examining the impacts of neighborhood design and residential self-selection on active travel: a methodological assessment. *Urban Geography*, 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. *Transportation Research Procedia*, *11*, 154-164.

Childers, T. L., Houston, M. J. & Heckler, S. E. (1985). Measurement of Individual Differences in Visual Versus Verbal Information Processing. *Journal of Consumer Research*, *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. *Transportation Research Part E: Logistics and Transportation Review*, *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, Cham, Springer International Publishing.

Economist, T. (2022). How the pandemic has changed American homebuyers' preferences. The Economist. UK.

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.").

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*: MIT press).

Greene, W. H. & Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, *37(8)*, 681-698.

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 Recognition*, *77*, 354-377.

Guevara, C. & Ben-Akiva, M. (2006). Endogeneity in Residential Location Choice Models. *Transportation Research Record: Journal of the Transportation Research 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. *Transportation Research Part B: Methodological*, *123*, 224-239.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C. & Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, *34*, 15908-15919.

Hanibuchi, T., Nakaya, T. & Inoue, S. (2019). Virtual audits of streetscapes by crowdworkers. *Health & place*, *59*, 102203.

Henik, A. & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & cognition*, *10*, 389-395.

Hess, S. (2014). Latent class structures: taste heterogeneity and beyond. In (Eds.), *Handbook of choice modelling* (pp. 311-330). Edward Elgar Publishing.

Hess, S., Adler, T. & Polak, J. W. (2007). Modelling airport and airline choice behaviour with the use of stated preference survey data. *Transportation Research Part E: Logistics and Transportation Review*, *43(3)*, 221-233.

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. *Theory and Decision*, *84(2)*, 181-204.

Hess, S. & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, *32*, 100170.

Hevia-Koch, P. & Ladenburg, J. (2019). Where should wind energy be located? A review of preferences and visualisation approaches for wind turbine locations. *Energy Research & Social Science*, *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.

Ito, K. & Biljecki, F. (2021). Assessing bikeability with street view imagery and computer vision. *Transportation research part C: emerging technologies*, *132*, 103371.

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.

Lee, J. & Huang, Y. (2022). Covid-19 impact on US housing markets: evidence from spatial regression models. *Spatial Economic Analysis*, *17(3)*, 395-415.

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*. (Berkeley: Institute of Urban & Regional Development, University of California).

McFadden, D. (2000). Disaggregate Behavioral Travel Demand's RUM Side, A 30 year retrospective. *In Travel behavior Research: The leading*, 17 - 64.

McFadden, D. L. (2001). Economic Choices. *The American Economic Review*, *91(3)*, 351-378.

Nisbett, R. E. & Wilson, T. D. (1977). Telling More Than We Can Know - Verbal Reports on Mental Processes. *Psychological Review*, *84(3)*, 231-259.

Parady, G., Ory, D. & Walker, J. (2020). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. *Journal of Choice Modelling*, 100257.

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. *Advances in neural information processing systems*, *32*.

Perez-Lopez, J.-B., Novales, M. & Orro, A. (2022). Spatially correlated nested logit model for spatial location choice. *Transportation Research Part B: Methodological*, *161*, 1-12.

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 intelligence and the future of testing*, *73*, 126.

Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, *2(1)*, 65-100.

922 Ramírez, T., Hurtubia, R., Lobel, H. & Rossetti, T. (2021). Measuring heterogeneous perception of
923   urban space with massive data and machine learning: An application to safety. *Landscape and*
924   *Urban Planning*, *208*, 104002.
925 Rossetti, T., Lobel, H., Rocco, V. & Hurtubia, R. (2019). Explaining subjective perceptions of public
926   spaces as a function of the built environment: A massive data approach. *Landscape and Urban*
927   *Planning*, *181*, 169-178.
928 Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla,
929   A. & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International*
930   *journal of computer vision*, *115*, 211-252.
931 Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M. & Villalobos, P. (2022). Compute trends
932   across three eras of machine learning. 2022 International Joint Conference on Neural Networks
933   (IJCNN), IEEE.
934 Sifringer, B., Lurkin, V. & Alahi, A. (2020). Enhancing discrete choice models with representation
935   learning. *Transportation Research Part B: Methodological*, *140*, 236-261.
936 Small, K. A. (2012). Valuation of travel time. *Economics of Transportation*, *1(1)*, 2-14.
937 Smith, B. & Olaru, D. (2013). Lifecycle stages and residential location choice in the presence of latent
938   preference heterogeneity. *Environment and Planning A*, *45(10)*, 2495-2514.
939 Stuart-Fox, M., Kleinepier, T., Ligthart, D. & Blijie, B. (2022) *Wonen langs de meetlat*. Ministerie van
940   Binnenlandse Zaken en Koninkrijksrelaties, DEN HAAG.
941 Szép, T., van Cranenburgh, S. & Chorus, C. (2023). Moral rhetoric in discrete choice models: a Natural
942   Language Processing approach. *Quality & Quantity*.
943 Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner,
944   A., Keysers, D. & Uszkoreit, J. (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances*
945   *in neural information processing systems*, *34*, 24261-24272.
946 Ton, D., Duives, D. C., Cats, O., Hoogendoorn-Lanser, S. & Hoogendoorn, S. P. (2019). Cycling or
947   walking? Determinants of mode choice in the Netherlands. *Transportation research part A:*
948   *policy and practice*, *123*, 7-23.
949 Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. (2021). Training data-
950   efficient image transformers & distillation through attention. International Conference on
951   Machine Learning, PMLR.
952 Train, K. & Wilson, W. W. (2008). Estimation on stated-preference experiments constructed from
953   revealed-preference choices. *Transportation Research Part B: Methodological*, *42(3)*, 191-
954   203.
955 Train, K. E. (2003). *Discrete choice methods with simulation*. (New York: Cambridge University
956   Press).
957 Van Cranenburgh, S., Chorus, C. G. & Van Wee, B. (2014). Vacation behaviour under high travel cost
958   conditions – A stated preference of revealed preference approach. *Tourism Management*, *43(0)*,
959   105-118.
960 van Cranenburgh, S., Wang, S., Vij, A., Pereira, F. & Walker, J. (2021). Choice modelling in the age
961   of machine learning-discussion paper. *Journal of Choice Modelling*, 100340.
962 Van Wee, B. (2009). Self- Selection: A Key to a Better Understanding of Location Choices, Travel
963   Behaviour and Transport Externalities? *Transport Reviews*, *29(3)*, 279-292.
964 Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. (2022). Scaling vision transformers. Proceedings of
965   the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
966