Landscape and Urban Planning 202 (2020) 103887

Contents lists available at ScienceDirect

Landscape and Urban Planning

journal homepage: www.elsevier.com/locate/landurbplan

Using convolutional autoencoders to extract visual features of leisure and retail environments

Sam Comber*, Daniel Arribas-Bel, Alex Singleton, Les Dolega

Department of Geography & Planning, University of Liverpool, Roxby Building, Liverpool L69 7ZT, United Kingdom

A B S T R A C T

Visual characteristics of leisure and retail environments provide sensory cues that can influence how consumers experience and behave within these spaces. In this paper, we provide a computational method that summarises the "visual features" of shopping districts by analysing a national database of geocoded store frontage images. While the traditional focus of social scientific research explores how drivers such as proximity to shopping environments factor into location choice decisions, the visual characteristics that describe the enclosing urban area are often neglected. This is despite the assumption consumers translate visual appearance of a retail area into a judgement of its functional utility which mediates consumer behaviour, patronage intention and the image a retail location projects to passers-by. Such judgements allow consumers to draw fine distinctions when evaluating between competing destinations. Our approach introduces a deep learning model known as Convolutional Autoencoders to extract visual features from storefront images of leisure and retail amenities. These features are partitioned into five clusters before several measures describing the environment around the leisure and retail properties are introduced to differentiate between the clusters and assess which variables are distinctive for particular groupings. Our empirical strategy unpacks different groupings from the clusters, which implies the existence of relationships between visual features of shopping areas and functional characteristics of the surrounding urban environment. Ultimately, using the example of retail landscapes, the core contribution of this paper demonstrates the utility of unsupervised deep learning methods to research questions in urban planning.

1. Introduction

Visual characteristics of urban spaces drive how individuals evaluate and experience their surroundings for the purpose of location choice behaviour and patronage decisions (Hauser & Koppelman, 1979). In the *The Image of the City*, Kevin Lynch argues the built environment can be drawn as "mental maps" that describe how the city is read visually by cues such as shapes, sizes and colours (Lynch, 1960). Not only this, Silver and Clark (2016) argue the actions, tastes, and traits of individuals create and support particular meanings attached to places. The measurement of a scene assesses the character of a particular place and highlights distinctive visual aspects of the built environment. As visual (but subjective) measures that describe scenes such as liveliness are hard to quantify with traditionally-available data, urban planners typically resort to building indicators that are based on more directly observable characteristics such as population density (Glaeser and Gottlieb, 2009) or street layout (Jung et al., 2017). Often representations that characterise the scenes of streets are inferred using visual audits conducted by researchers who collate data to explore similarities and differences of physical attributes visible from street-level – the quality of building facade, the presence of street art, or the condition of sidewalks, for example (Bader et al., 2017). Once aggregated, researchers can unpack relationships exploring the link between particular visual attributes of built environments and characteristics of the surrounding area. For retail environments, the visual image that shopping areas project to consumers is a function of a broad range of influences which affect patronage behaviour and consumer experiences (Bell, 1999). Retail area image is a multidimensional concept and to understand it requires unpacking the multitude of functional and visual characteristics that consumers associate with shopping areas (Baker et al., 1994). These characteristics are stimuli that influence consumer perception and, by extension, patronage intention for particular retail environments. Typically measures of retail area image are derived using survey approaches that rate characteristics such as the quality of building materials, the attractiveness of shop signage and overall environmental cleanliness (Bellizzi et al., 1983; El-Adly, 2007). As conducting in-person studies in shopping areas to record this data requires a high level of human judgement, they are cost-intensive and limited in the throughput required to construct visual descriptors of retail environments for large study areas

To circumvent the scalability issues of manually auditing a national sample of retail locations, we apply Convolutional Autoencoders (CAEs) to automatically extract visual features from images showing the frontage of leisure and retail properties across England and Wales. Particular interest on street-level imagery for leisure and retail amenities stems from their influence to the vibrancy of places and, hence, in

* Corresponding author.
E-mail addresses: s.comber@liverpool.ac.uk (S. Comber), D.Arribas-Bel@liverpool.ac.uk (D. Arribas-Bel), Alex.Singleton@liverpool.ac.uk (A. Singleton), L.Dolega@liverpool.ac.uk (L. Dolega).

https://doi.org/10.1016/j.landurbplan.2020.103887
Received 8 November 2019; Received in revised form 15 June 2020; Accepted 21 June 2020
Available online 11 July 2020
0169-2046/ © 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

the characteristics of the urban hierarchy (Dennis et al., 2002). While previous studies have shown that proximity to leisure and retail amenities factor into location choice decisions and patronage of retail environments (Glaeser and Gottlieb, 2009), the visual characteristics that describe the urban environment around the point of interest are neglected. Such approaches assume a "vacuum" around single amenities, which ignores the environmental context that surrounds these premises. As an example, the visual characteristics of a street with a restaurant accessible by several modes of transportation is likely to differ by the amount of liveliness when compared with another restaurant serviced in a location with no transport links. Capturing visual features of leisure and retail amenities allows an exploration into whether aspects of *what we see* are related to particular characteristics of the built environment that describe the amenities location. By clustering visual features extracted from the CAE, the principal contribution of this paper uses deep learning to assess whether visual-only features of retail landscapes correlate with observed characteristics of built environments, and whether there are distinctive characteristics for particular groupings.

The remainder of the paper is organised as follows. Section 2 motivates the underlying conceptual framework of the paper. Section 3 introduces the sources of data we utilise through the study, before describing the modelling approach we implement to arrive at our empirical objective. Section 4 presents the main findings. Finally, Section 5 concludes the paper.

2. Background and motivation

2.1. Visual characteristics of built environments

In the *Critique of Judgement* Immanuel Kant first observed aesthetic perception as a self-organising process that drives how individuals react to different environments (Kant, 1790). Not only do humans perceive their environment as neutral facts and data, but we react to distinctive aesthetic cues encoded in our surroundings that change how these spaces are experienced as we walk through them (Silver and Clark, 2016). Our judgement of the elements in our surroundings are rendered as a totality, independent of the constituent parts. When we stroll through a "hip neighbourhood", the avant-garde feel, boutique stores, and DIY atmosphere are not perceived as independent objects. This is because they collectively recall a particular way of behaving that is adopted from the tastes and preferences derived from the environment the individual chooses to surround themselves with (Merleau-Ponty, 2004). Thus, an environmental psychology influences how preferences for certain environments are driven by a multitude of interwoven factors. Jane Jacobs recognised this as early as 1960, emphasising the role streets perform in setting the visual scene of cities. In a critique of modernist planning policy, Jacobs (1961) argued that unifying design elements of urban spaces is short-sighted, as the interplay of their "bits and pieces" are central to supporting the diverse excitement that street scenes offer.

Visual cues are seen as discriminative features that influence perceptions and evaluations of urban spaces, and even when considering socio-cultural biases in aesthetic judgement, have been shown to affect the psychological state of their inhabitants (Quercia et al., 2014). Kelling and Coles (1997)'s *Broken Windows Theory*, for example, suggests cues of environmental disorder in urban appearance such as abandoned cars, litter, and vandalism drive a perceived breakdown of social order which, in turn, induce more severe forms of criminal activity. Beyond disorder places deviate from conventional form by appearing, amongst other things, transgressive, glamorous, or informal (Silver and Clark, 2016). Thus, a suite of evaluative dimensions are considered when characterising the visual attributes of urban spaces, with different environments reflecting different visual representations of tastes and values. Not only this, Massey (1991) argues these particular spaces are not static, but have multiple identities that are forged by ever-changing social interactions occurring between people within them. All together, these considerations highlight the complexities of capturing a signal that reflects the visual qualities of street scenes.

2.2. Traditional approaches for describing retail environments

As aesthetic descriptions of urban environments such as glamorous, lively or conventional are difficult to measure directly, urban scientists typically fall back to constructing indicators of the qualities that describe spaces such as shopping areas (Silver and Clark, 2016). In-person visual audits strive to unpack how the functional, physical and social characteristics of retail environments correlate to affective outcomes such as store patronage and location choice decisions. Survey techniques have an extensive history in urban planning research, and borrow from psychometric measurement models to infer latent traits through an aggregation of single items visible across the audit (Bader et al., 2017). In UK planning discourse, for example, concepts such as vitality and viability have long underlined 'health checks' of town centre areas, reflecting arguments in Jacobs (1961) that thriving places maintain a diverse range of uses, attract significant numbers of people, and sustain a continuing ability to attract investment (Ravenscroft, 2000). Thus, vitality and viability is typically inferred by aggregating multiple items such as pedestrian counts, diversity of amenities, or boarded-up windows that are sampled at points across different retail locations. In the retail literature, several examples aggregate sets of measures to describe visual characteristics of shopping spaces. Bell (1999), for example, shows environmental stimuli such as appealing store colours, attractive shop signs and fashionable product ranges constitute a 'visual amenity' that inspires consumer willingness to patronise a shopping environment. Moreover, El-Adly (2007), finds attractiveness attributes of shopping malls such as luxury, comfort and convenience drive different patronage motives amongst different shopper segments in UAE.

Survey-based approaches are often required to describe the visual properties of urban environments due to the absence of accessible and high coverage quantitative data (Salesses et al., 2013). Traditionally, studies are undertaken by relying on a mix of personal interviews, street-level observations of visual appearances, and annotated video recordings by experts (Quercia et al., 2014). This manual review of material is an arduous task however, and requires considerable collective effort to distinguish amongst the variety of visual cues encoded in the images.

2.3. Deep learning approaches for describing urban environments

To evaluate visual characteristics of particular places, Convolutional Neural Networks (CNNs) that are 'trained' with human-labelled images of street scenes are increasingly used to automate the classification of the scenes presented by built environments. This new body of literature has been punctuated by emerging access to new sources of data that have been released by commercial providers and photo-sharing websites in open formats (Arribas-Bel, 2014). Providers such as Google Street View (GSV) and Flickr have opened up access to street-level imagery for researchers through Application Programming Interfaces (APIs), which have, in turn, been used to construct modern crowdsourcing platforms for collecting millions of user perceptions about particular places. Large quantities of human-labelled, street-level imagery have been used for training computer vision techniques. Zhang et al. (2018), for example, use a deep learning based approach to predict perceptions of neighbourhoods in Bejing, China along six perceptual indicators of safe, lively, boring, wealthy, depressing, and beautiful, before investigating which visual elements correlate to a particular perception. The study used street-level images collated by MIT Media Lab as part of the "Place Pulse" program, which by fall 2018 had collected 1,566,218 pairwise comparisons between 110,988 street-level images from 56 cities worldwide (Dubey et al., 2016). This crowdsourced data was made publicly available by Salesses et al. (2013), who originally used it to understand the effect of the built

environment's visual features on perceptions of safety, class and uniqueness in the cities of New York and Boston in the United States, and Linz and Salzburg in Austria. Additional studies that use labelled GSV images include Liu et al., (2016), who detect shifts in city identities and urban form for 26 cities from Europe, Asia, and North America. Seresinhe et al., (2017) trained machine learning models on 217,000 crowdsourced images from the "Scenic-Or-Not" online game that rates outdoor, natural environments on an integer scale (1–10) of its *scenicness*, and explores questions that ask which types of greenspaces are perceived as beautiful.

Unfortunately, a drawback of these supervised methods are the large sample sizes required to train the network which are often unfulfilled in real-life applications. Moreover, these approaches typically utilise a large, non-expert workforce (voting on crowdsourcing platforms) to construct massive volumes of labelled image data. This creates several challenges. Principal amongst these is the balancing between maintaining a swift and economical annotation process while ensuring the collected labels are accurate (Sorokin and Forsyth, 2008). More importantly, the user's interaction with the labelling task may be influenced by socio-economic and demographic factors. As urban experiences are highly socially constructed, different groups might engage with the built environment in different ways, meaning visual characteristics are highly particular to various socio-economic or demographic groups (Quercia et al., 2014). These challenges exist because CNNs are supervised, meaning they require the network to be shown labelled instances of images for learning the nuances between particular predicted outputs. An alternative approach to extracting features from street-level imagery are *Convolutional Autoencoders* (CAEs). CAEs are *unsupervised* approaches meaning they provide a self-organised means for learning the relationships between elements in the data without being shown labelled inputs. CAEs are advantageous because they provide a less data-intensive alternative to CNNs that does not require the user to assemble large quantities of labelled data for training the network.

2.4. Application of computer vision methods to retail environments

While many studies that apply deep learning have focussed on urban environments, to our best knowledge, no application of deep learning to explore visual characteristics of retail environments currently exists in the literature. This is despite the high suitability of computer vision methods for characterising the variance in image attributes between different shopping areas. Consumers with little experience of a store or environment may use perceptual qualifications of image, in addition to prices, as a proxy for the quality of goods and service provision (Bell, 1999). Stimuli that influence consumer perceptions of shopping area image are functional qualities but also the aura of psychological attributes aroused by the environment. Functional characteristics include convenience and accessibility of store or retail area location, parking availability, the range of stores and products offered, and proximity to residential neighbourhoods and workplaces (Baker et al., 1994; Chebat et al., 2010). Psychological characteristics relate to the "visual amenity" experienced by consumers in shopping environments. For example, previous research links store patronage decisions to visual elements such as architecture, shop signage and exterior design (Baker et al., 1994), but also factors such as cleanliness and even colour of store premises (Bellizzi et al., 1983). Thus, quality inference for shopping areas is a function of multiple influences that affect consumer decision-making choices.

Given the wealth of research that has already linked image attributes of shopping areas to consumer patronage, the focus of the present study moves away from an exploration of footfall. Instead, our main research direction focuses on characterising the different *visual* representations of shopping environments by functional attributes that describe the area in which the premise is located. In synthesis of these two attributes, we unpack different representations of the *scene* that particular shopping environments project to passers by. The "*scene*" of an environment reflects both the visual characteristics and configuration of leisure, services, retail and cultural life, and data describing amenities such as leisure and retail premises are windows that allow researchers to unpack these configurations (Silver and Clark, 2016). An understanding of different *scenes* from leisure and retail environments is an important exercise because it unpacks patterns of urban human activity and function. This is useful information for retail planners and urban management schemes because it raises awareness of attributes and image among particular areas. Public or private sector agencies might utilise this to rationalise investment decisions that allocate spend to promotional activities and place marketing campaigns for building the profile of shopping environments (Page and Hardyman, 1996).

The visual design of retail environments are among the tools used to enrich the consumer shopping experience. Visual design of shopping areas has been manipulated previously to evoke desirable responses, such as arousal and pleasure which triggers approach behaviour and supports store positioning (Ballantine et al., 2010; Baker et al., 1994). Yet the visual design of retail environments in the UK is highly particular, and so consideration to its nuances is required for understanding potential implications to our applied methods. One limitation in applying computer vision methods to UK high street environments is a phenomena known as *clone towns* (Ryan-Collins et al., 2010). The idea of 'cloned' streets relates to the loss of identity and local character when chain stores come to homogenise high street environments at the expense of independent stores (Carmona, 2015). The implications for computer vision approaches concern the difficulty in identifying different typologies where no unique characteristics are directly observable from the images when they broadcast no local distinctiveness. Despite the British Retail Consortium (2009) arguing there have been calls for communities to reclaim their local high streets through the encouragement of local spending, it remains that a large number of distinctive facades constructed from local building materials may have been exchanged by identical glass, steel and concrete frontages (Ryan-Collins et al., 2010). This potentially limits the discovery of more interesting, diverse and distinctive types derived from empirical exercises that use street-level imagery from UK high streets. Despite this limitation, for wider study areas than would be permitted by in-person audits, computer vision approaches allow us to unpack how visual features of leisure and retail properties relate to functional characteristics of shopping environments, and consequently, how we can characterise the scenes these places offer.

3. Empirical strategy

Our approach to explore differentiation between visual features of leisure and retail premises is three-staged. Firstly, we extract visual features from images of leisure and retail premises using a computer vision algorithm. Secondly, we partition visual features into a sensible number of clusters using a bottom-up classification strategy. And thirdly, to differentiate between the clusters, we introduce variables that describe characteristics derived from the point of interest around the properties.

3.1. Data

To implement the methodological approach we require two principal sources of data described below. Our first source of data are street-level imagery of 314,542 retail, service and leisure properties across England and Wales. These images display the front exterior of the property that face onto the adjacent street or open space. Exterior images were collected by a large pool of surveying teams equipped with hand-held cameras from the Local Data Company (LDC) in 2015. Sample images are displayed in Fig. 3.1, and are categorised row-wise by several variables introduced in Table 3.1. As a pre-processing step, each JPEG image is resized from $800 \times 400 \times 3$ to a $224 \times 224 \times 3$ pixel

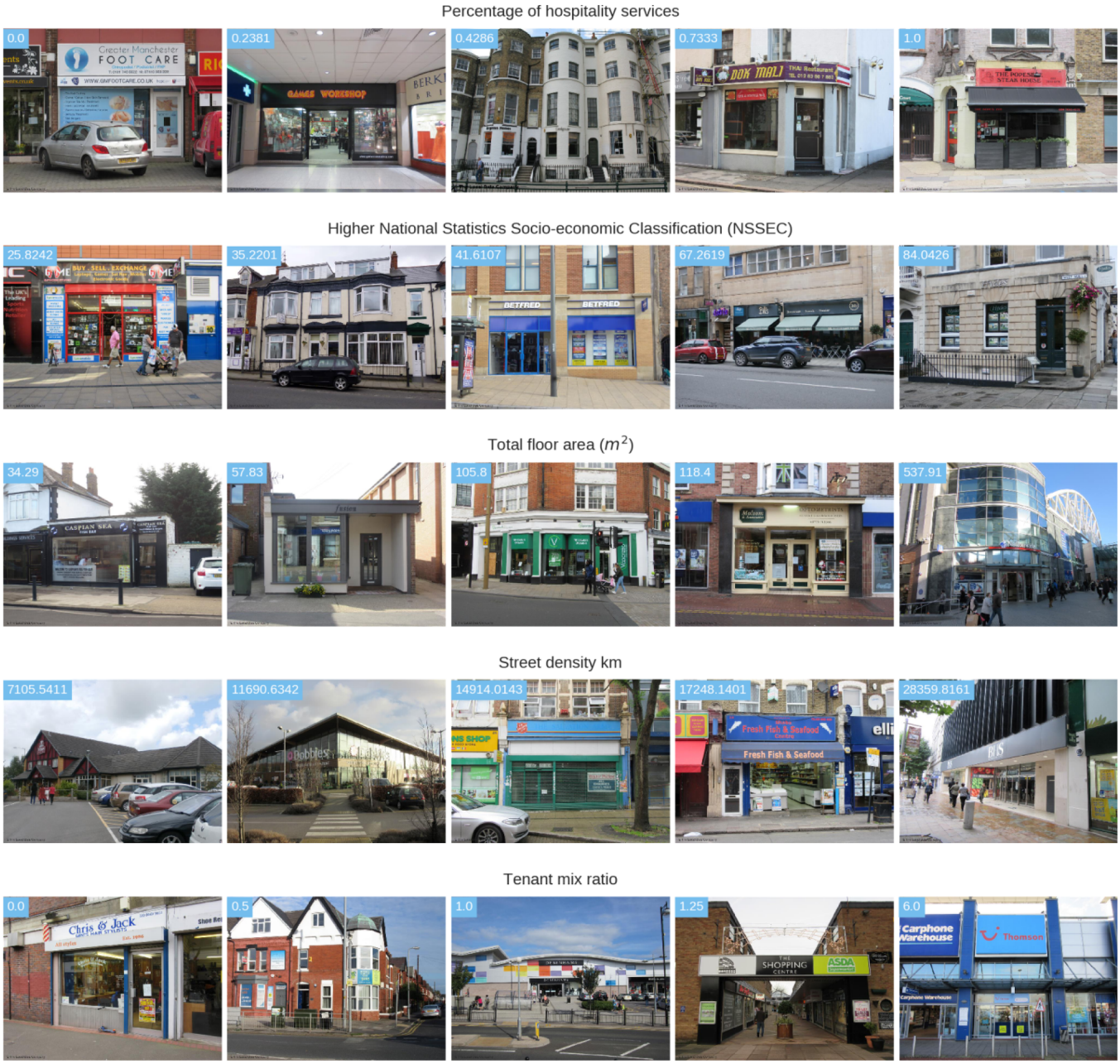

Fig. 3.1. Sample LDC images for several features in Table 3.1. Each image is a random sample from each of five equal interval bins.

image for compatibility with the applied neural network architectures, before normalizing the RGB values (0–255) to a 0–1 range. These resized, normalised digital images are the 3-dimensional inputs (width, height, and colour channel) to the convolutional neural networks we introduce in Section 3.2.

While this data offers new opportunities, there are limitations of using street-level imagery for visual audit purposes. Channels that affect perceptions of built environments such as sound and smell are absent from pictographic representations, and so cannot be directly evaluated from the image (Salesses et al., 2013). Similarly, small items less visible to the human eye that vary over short periods such as litter, drug paraphernalia, broken glass, or cracked sidewalks are difficult to measure given street-level imagery represent a single snapshot in time (Bader et al., 2017). More specifically, given the principle concern for the LDC surveying teams was to photograph facade features of the store premises, measures related to sidewalks such as number of parked cars or shrubbery might be partially occluded in the image, despite contributing to the overall ambiance of the urban area. Despite these limitations, the LDC images remain a valid source of data for our purposes. This is because they simulate a virtual walk down the street that replicates an eye-level experience, and the large number of LDC images provides granular, unprecedented coverage that would be impractical (and cost-intensive) to obtain otherwise.

The second source of data is derived from characteristics that differentiate the particular visual representations of LDC images, and is used in the third stage of our approach. Our variable selection covers measures derived within a 15-minute walk catchment (assuming a walk speed of 4.5 km per hour) around each leisure and retail premise (see Fig. 3.2). These catchments are constructed using OSMnx, which is a Python library for acquiring, analysing and visualising street networks (Boeing, 2017). Within each catchment, we derive measures for a number domains outlined in Dolega et al. (2019) that describe shopping activity such as composition, diversity, size and function, and economic health (see Table 3.1). Aside from LDC and OSMnx data, we derive

Table 3.1
Variable description for the domains of economic health, composition, size and function and socio-economics of leisure and retail premises.

Variable	Description	Source	Mean	Std. Dev	Unit
Economic health					
bus_rate	Rateable value taxed on the business property.	LDC	100,639.9	976,771.5	Pounds
vac_rate	Vacancy rate of Local Authority District the property resides in.	LDC	0.09	0.03	Percent
unemployed	Percent of unemployed people in Output Area.	ONS	5.75	3.64	Percent
e-res_score	E-resilience score of nearest town centre.	CDRC	0.08	0.45	Score
transport	Number of bus or train links within catchment	NaPTAN	61.21	38.19	Count
Composition					
comparison	Proportion of comparison goods stores within catchment (clothing, household goods, etc).	LDC	0.21	0.21	%
hospitality	Proportion of hospitality outlets within catchment (restaurants, bars, etc).	LDC	0.31	0.24	%
convenience	Proportion of food retailers within catchment (grocers, butchers etc).	LDC	0.13	0.18	%
consumer	Proportion of consumer services within catchment (banks, estate agents, etc).	LDC	0.18	0.21	%
tenant_mix	Retail to service ratio of catchment.	LDC	0.88	1.00	Ratio
store_diversity	Diversity of store types within catchment calculated by Shannon entropy.	LDC	1.14	0.57	Bit
Size and function					
floor_area	Total floor area for the property.	LDC	227.61	830.08	m^2
car_parking_spaces	Number of car parking spaces at the property.	LDC	1.28	17.36	Count
roeck_compactness	Compactness of catchment morphology.	OSMnx	0.49	0.13	Ratio
store_diversity	Number of stores within catchment.	LDC	14.24	20.45	Count
eig_centrality	Influence of store location within street network of catchment.	OSMnx	0.02	0.04	Score
street_length_avg	Average length of streets in catchment.	OSMnx	66.43	25.86	Meter
street_density	Total street length within catchment divided by catchment area.	OSMnx	15911.72	7134.42	km^2
Socio-economic					
high_nssec	Percent of people with higher occupational employment in Output Area.	ONS	42.86	17.21	Percent
detached	Percent of housing units classified as detached in Output Area.	ONS	6.09	9.52	Percent
flats	Percent of housing units classified as flats in Output Area.	ONS	35.13	24.94	Percent

variables from several other sources. Census data is provided by the (ONS, 2016), our *e-res_score* variable is from a Consumer Data Research Centre (CDRC) data product and describes the vulnerability of town centres to the impacts of online shopping (estimated by Singleton et al., (2016)), and the *transport* variable is from the database of National Public Transport Access Nodes (NapTAN) (Department for Transport, 2014). In addition, we use a small number of census-based socio-economic characteristics at Output Area (OA) level to describe the area in which the leisure or retail premise resides. OAs are built from postcode units and are the smallest statistical unit for which UK census data is published (ONS, 2019).

3.2. Visual features from CAEs

Given the collection of leisure and retail property images are unlabelled and represented by a large number of raw pixels, a mathematical technique is required to decompose this larger set of correlated variables (or pixels) to a condensed set that captures the most salient characteristics of the image (Efron and Hastie, 2016). To learn this compressed set of variables from the raw pixels we rely on Convolutional Autoencoders (CAEs) (Goodfellow et al., 2016) which are composed of two layers: an encoder layer f_E and a decoder layer f_D. From a non-technical standpoint, the objective of CAEs is to take an input image, I, and reconstruct it as a copy, $\hat{I}$. Internally, CAEs use a hidden layer h that describes a code to reconstruct the image (Goodfellow et al., 2016). This lower dimensional mapping forces the CAE to prioritise aspects of the image that are the most useful for reconstructing a copy from the input image, meaning h learns the most useful properties of the data while discarding redundancies.

CAEs are extensions of autoencoders, which are techniques that essentially reduce the data under consideration to a smaller set of principal values. Practical applications of autoencoders include data compression for saving storage space and transmission times, and also cleaning corrupted data inputs by denoising. Thus, CAEs are autoencoders that introduce convolutional and (de)convolutional layers in the encoder f_E and decoder f_D sections, respectively:

$$f_E = \sigma(I * K + b) = h$$

where σ is a Rectified Linear Unit (Relu) activation function which is a truncation performed individually for every pixel x of the input, $Relu(x_{ij}) = max(0, x_{ij})$, that allows the CAE to learn non-linear patterns in the data, I are $224 \times 224 \times 3$ images where the 3 refers to the red, blue and green (RGB) colour channels, K are 3×3 matrices called convolutional filters, b is the bias unit which is similar to the intercept of a linear function and allows the line of the activation function to shift from the origin, and h is the code that represents the lower dimensional mapping of I. The convolution operator, $I * K$, is described more explicitly for the first layer in Eq. 3.2:

$$(I * K)_{xy} = \sum_{i=1}^{224} \sum_{k=1}^{224} K_{ij} \cdot I_{x+i-1, y+j-1}$$

which overlays each 3×3 filter over every possible pixel of the image, and records the sum of the element-wise product to an intermediate representation known as an activation map. The convolutional operator exploits spatial location in the image, as neighbouring pixels become activated for particular groups of edges that respond to semantically meaningful objects – trees, cars, or people, for example. This means particular filters become activated for specific patterns in the image, and stacking these filters across successive convolutional layers facilitates *parameter sharing*, where hierarchies of filters introduce levels of abstraction to the different kinds of features identified in the image (Goodfellow et al., 2016). As an example, the banks of filters learnt at the first convolutional layer might represent lower-level features such as lines, circles, and curves, while the higher-level convolutional layers will use these to construct whole objects – eye-like shapes or automobile wheels, for example. As the starting values of the K filters are randomly initialized, over the course of training the CAE the network will learn to find the optimal filter values that minimize the reconstruction error between I and $\hat{I}$.

Within each convolutional layer, a final step commonly applied to modify the output from Eq. 3.2 is pooling. We apply the max *pooling* operation which returns the maximum pixel value within a 2×2 filter that steps across non-overlapping pixels of the input. This has the net effect of down-sampling an image by a factor of two, which sequentially reduces the pixel representation of our image from $224 \times 224 \times 3$ to a *latent* representation, h, which has shape $28 \times 28 \times 1$ and reflects the

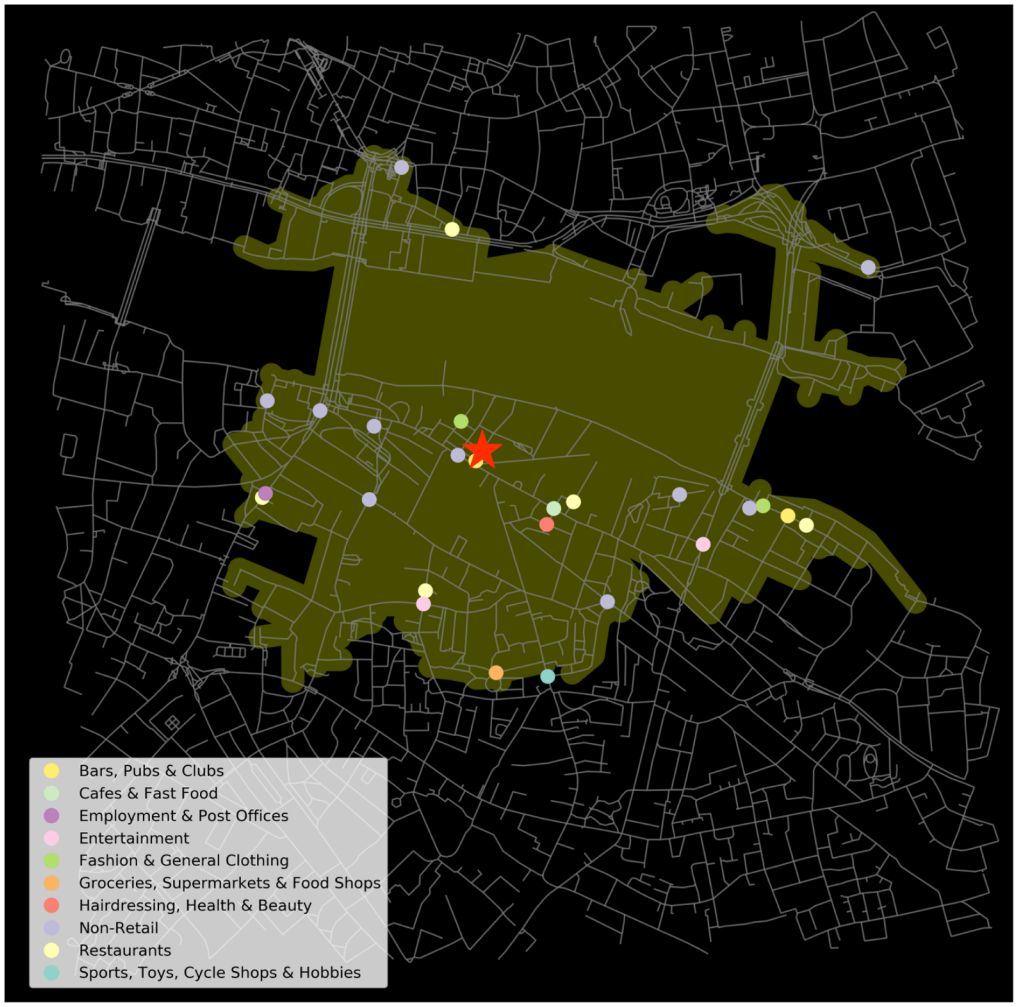

Fig. 3.2. Example 15-minute walk catchment for a retail store around London Bridge. Note: 30 leisure or retail premises are sampled within the catchment to avoid clutter. Large red star denotes the store for which the catchment was created. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

visual features we use for our clustering exercise (see Section 3.3).

To train the CAE end-to-end, we also require a decoder f_D network that reconstructs the original image $\hat{I}$ from h:

$$f_D = \sigma(h * U + b) = \hat{I}$$

The only difference between f_E and f_D is that convolutional layers in the former are replaced by deconvolutional layers in the latter. This has the net effect of up-sampling the latent representation h ($28 \times 28 \times 1$) back to $224 \times 224 \times 3$, thus completing the reconstruction of the original image I. Once the CAE network has been sufficiently trained, the latent representation h, represented by $28 \times 28 \times 1 = 784$ pixels, becomes the basis of the visual features we use to differentiate between the visual scenes of different leisure and retail premises. To summarise these methodological steps, we visualise the resulting CAE architecture defined by Eq. 3.1 and Eq. 3.2 in Fig. 3.3. In regards to implementation, the CAE model is defined in Keras (Chollet, 2015), with training undertaken on a single Nvidia Quadro M4000 GPU with 8 GB memory.

3.3. Clustering visual features

To derive meaning from the visual features, we require a technique to group our vectors of visual features such that those in the same grouping exhibit similarities. This allows us to unpack similarities between the visual scenes for different retail environments which we can then describe by a number of functional characteristics outlined in Table 3.1. Our approach constructs a *bottom-up* classification where an initial typology with 250 numerous smaller groups are partitioned using k-means. Given the sensitivity of k-means to the initial starting values of the centroids, the algorithm is initialized 1000 times with different centroid seeds, taking the final result as the output that best minimizes the within-cluster sum of squares. Finally, we allow up to 100,000 iterations within a single run to ensure stable convergence of the centroids. After the initial partition, we aggregate the clusters into coarser and larger groupings based Ward's method of hierarchical clustering (Ward, 1963). As Ward's method produces a dendrogram, we use it to slice a horizontal cut along the y-axis to create coarser levels of classification, which groups the 250 centroids of visual features into a smaller number of distinct clusters. This final partition represents the resulting clusters that differentiate the visual characteristics of the LDC images. Thus, we replicate a work flow similar to Spielman and Singleton (2015) and follow simple and widely supported methods to facilitate methodological transparency and reproducibility.

4. Results

In this section, we develop a discussion of our empirical findings based on two validation procedures. First, we undertake a validation exercise on our bottom-up clustering solution to ascertain a desirable

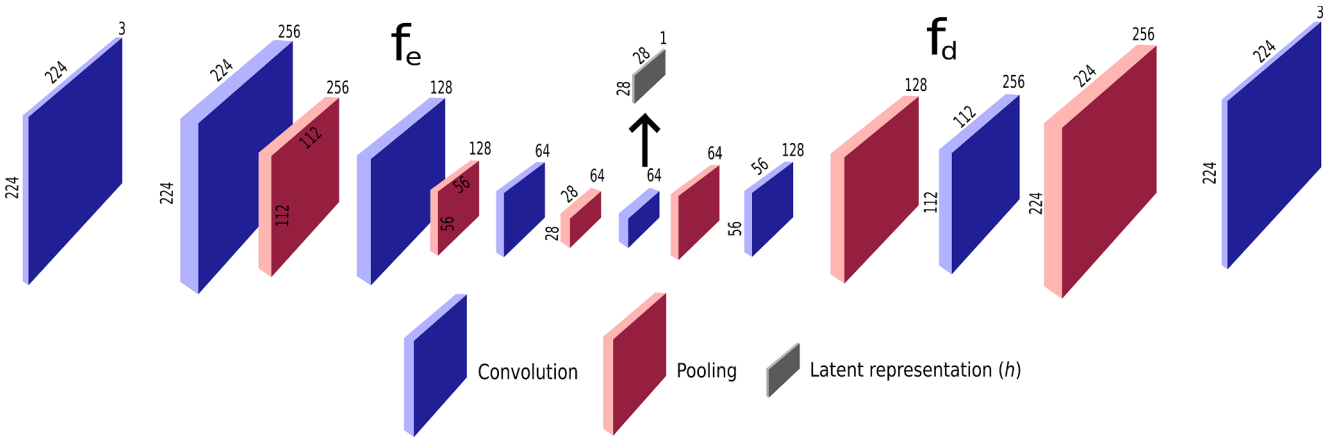

Fig. 3.3. Convolutional Autoencoder (CAE) architecture showing encoder f_e, compressed representation h, decoder f_d and reconstructed LDC image $\hat{I}$. Note: filter numbers are shown horizontally along z-axis of feature maps, while width and height are shown along the x and y, respectively. Illustration was produced on the open-source vector graphics editor Inkscape (Inkscape Project, 2019).

number of clusters; and second, we explore consistency of group membership to particular clusters across sets of visual features generated from the CAE and two pre-trained CNNs. For brevity, the detailed outcome of these exercises are moved to Appendix A and Appendix B. Based on the outcome of these exercises, in the following section we introduce several characteristics to unpack differences between the *five* distinct clusters of images we retrieve from our clustering approach.

4.1. Differentiating visual characteristics

To describe differences between the visual clusters, we aggregate characteristics for the consumer properties from Table 3.1, taking the median value for each variable per cluster. Prior to the aggregation, we transform each variable to z-scores by standardization, $z = \frac{x-\mu}{\sigma}$, meaning each characteristic is rescaled by the fractional number of standard deviations from the mean value. To begin, we introduce radar plots in Fig. 4.1 where each plot reflects a different visual cluster that shares similar psychological attributes reflected by common visual elements such as similar exterior design, signage, architecture, or colour. Along the axis of each plot aggregated variables that describe functional characteristics of these clusters are displayed. Thus, in synthesis of visual (psychological) attributes revealed by the cluster groupings and functional characteristics by the variables, we describe the *scene* projected by the clusters.

Turning to the group sizes, we note the numbers of leisure and retail premises within the visual clusters vary substantially. Our largest cluster, Group A, contains 159,251 leisure and retail properties whose built environment is distinguished by high density street networks and large proportions of comparison retail outlets who sell merchandise that consumers purchase relatively infrequently and so evaluate prices, features and quality between stores before making a purchase. This includes outlets such as DIY & household goods, electrical, and clothing and footwear stores. Group A also contains a considerable proportion of hospitality outlets such as restaurants, bars and pubs, and entertainment venues. The Roeck compactness value measures irregularity in the shape of the retail area's boundary, with higher values indicating a highly compact retail area and lower values reflecting dispersion. The Roeck value for Group A, alongside its high street density, implies the urban morphology of the built environment around these stores is highly dense and not dispersed. All together, this suggests the scene characteristics of Group A reflects a bustling shopping area with relatively affluent residents who live in the immediate area (as shown by the high percentage of residents in higher occupational roles).

Group B contains 24,567 leisure and retail premises and is highly differentiated amongst its characteristics when compared to the other clusters. The functional attributes shared by leisure and retail premises inside this visual grouping reflect areas that have a low diversity of premise types, with the majority of outlets represented by comparison retail or consumer services such as car showrooms and house & home stores. Premises in this cluster are located in areas with high vacancy rates, meaning there are higher percentages of vacant or unoccupied store units relative to the other groupings. Moreover, outlets in this cluster appear to have high total floor areas and are serviced by fewer transport options, which conjures images of peri-urban spaces consisting of large retail units and warehouse spaces located on the fringes of dense urban areas and so are less beaming with consumer activity. Overall, the visual and functional characteristics of Group B portray a scene of sparse and less desirable retail and leisure land use when compared with the other clusters. This is reinforced by socio-economic characteristics which reveal that individuals who live in the area, and might patron the shopping environment as consumers, typically occupy low percentages of high paid employment.

The next grouping that shares visual similarity is Group C, which contains 81,310 leisure and retail premises and is ascribed the label of 'Upmarket Hospitality'. The shopping environment of premises in this cluster are reflected by a large proportion of diverse hospitality outlets and leisure venues. This includes services ranging from restaurants and bars to theatres and galleries. A second defining characteristic of Group C is the extremely low vacancy rate when compared with the other clusters. This shows store units around the built environment for this grouping are typically occupied, which implies units in this cluster are in higher demand and so possibly elicit increased rates of rent. Similar to Group A, catchments around premises in this cluster are well served by transport links and possess highly similar urban morphology and socio-economic characteristics. In synthesis of visual similarities for leisure and retail premises within the cluster and functional characteristics of the urban landscape around these premises, Group C projects the scene of a thriving and upmarket shopping environment that is highly accessible and amenable to consumption activity.

Our smallest grouping, Group D, contains 6,962 leisure and retail units and is highly similar to Group C, although there are a few variables that differentiate the two clusters. Like Group C, Group D is characterised by a diverse range of hospitality outlets and stores that provide comparison goods such as electrical appliances and clothing. Compared to the dense street network of Group C, the urban morphology of Group D appears to reflect longer average street lengths that

Fig. 4.1. Median economic health, composition, size and function, and socio-economic characteristics in standardized units. Circular red line identifies zero, which shows standard deviations from the mean value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are fairly dispersed as shown by the low street density. Consistent with conventional wisdom, these two observations imply the built environment surrounding leisure and retail premises of Group D reflects high street shopping areas. Residents who occupy residential housing near stores in Group D typically occupy lower proportions of higher managerial roles. This suggests consumers, and by extension local consumption opportunities, are represented by less upmarket leisure and retail outlets given local patrons are typically less affluent than in Group C. Nine example images comparing low to high average street length for Group A and D, respectively, are shown by Fig. 4.2. The presence of automobiles in images sampled from Group D suggest the built environment here is more amenable to vehicle use, with streets around leisure and retail premises in this cluster typically longer and less dense. All together, the composite visual and functional characteristics of Group D project a scene of long high streets that serve a diverse range of consumption purposes to local consumers.

The last cluster, Group E, contains 81,310 leisure and retail premises and represents a middle ground between Group C and Group E. While units providing hospitality represent the highest proportion of services in this cluster, no particular mode of retail or leisure dominates unlike the other groupings. In fact, premises in Group E have the lowest proportion of comparison retailers in the surrounding urban environment. The urban morphology of Group E is fairly dense and compact, as evidenced by a relatively high street network density and Roeck compactness value. In synthesis, the shared functional attributes of premises in Group E suggest this grouping reflects a leisure, services and shopping environment that is accessed by consumers for everyday consumption as opposed to being accessed for a particular mode of retail or leisure service.

5. Discussion and conclusions

Visual characteristics of shopping environments are a significant determinant of area consideration and choice (Bell, 1999). Traditionally, visual representations of retail areas are retrieved using teams of human surveyors, who are cost-intensive to train and limited in the throughput necessary to construct the visual form of built environments. Consequently, in this paper, we use vast quantities of street-level imagery to explore whether visual features of leisure and retail environments correlate to measurable characteristics of built environments. This was achieved using a deep learning model known as Convolutional Autoencoders (CAEs) which learnt a compressed representation that captured the most salient characteristics required to reconstruct the image from a lower dimensional representation. Once these visual features were partitioned into a sensible number of clusters, functional characteristics that describe a 15-minute walk catchment from each premise were introduced to differentiate between the cluster partitions. By clustering the compressed representation, we were able to identify five partitions from the data that reflected different categorisations of the *scene* that particular shopping environments project to consumers across a national extent. This is important because information describing retail area image has historically been desired by retail planners for rationalising investment decisions in place marketing campaigns (Page and Hardyman, 1996), but is seldom available at wide geographical scales.

Furthermore, our findings unpacked patterns of retail activity and function, which demonstrated that certain visual features were distinctive for particular built environments. From an urban planning perspective, the main implications of our study demonstrated that

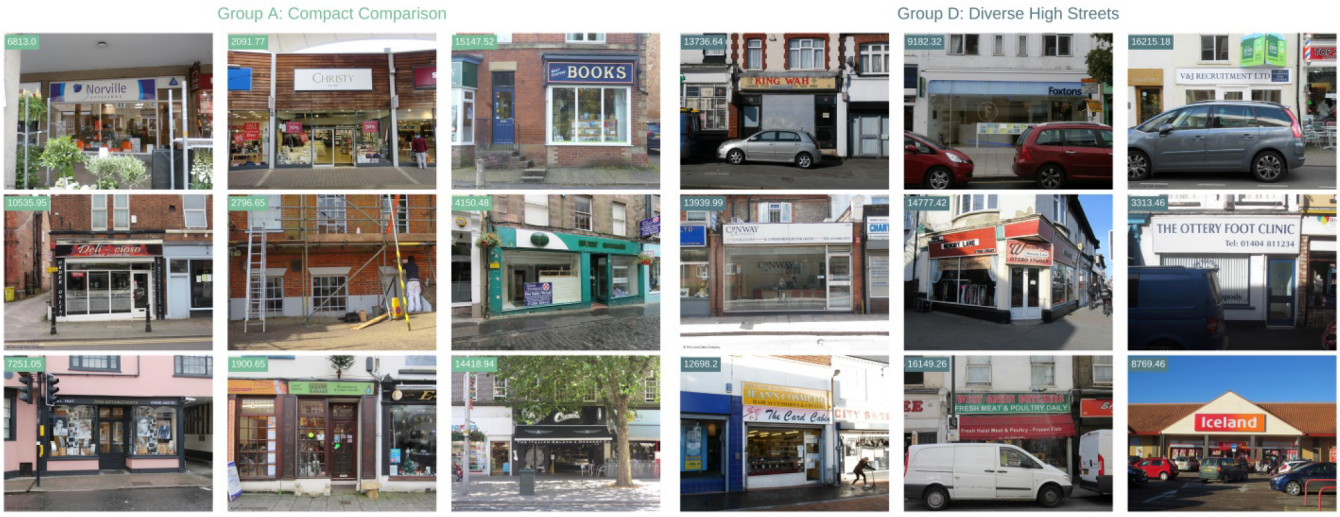

Fig. 4.2. Leisure and retail storefront images and average street length values in metres sampled from Group A and Group D.

aspects of what humans *see* were related to particular functional characteristics of retail environments. This was a pertinent question for retail practitioners to ask, as while previous studies have shown that *proximity* to (and *attractiveness* of) amenities such as leisure plazas, galleries and shops enter into consumer patronage decisions (Glaeser and Gottlieb, 2009), the defining visual characteristics of these environments are typically ignored. This is despite visual amenity being an important influence on patronage behaviour and the *scene* that shopping environments project to consumers (Silver and Clark, 2016).

In more practical terms, our approach could be mobilised within retail planning by adding a visual dimension to retail site optimization tools, and be used to optimally locate stores in locations suitable to particular consumer space uses. More precisely, retail managers could take photographs of prospective site locations, and classify each one according to the several clusters we identify. This would require passing the photographs through the CAE, and using the clustering outcomes fitted on the LDC images to predict cluster membership of these new, unseen photographs. Our approach, therefore, could be used to contextualise the visual qualities of potential store locations among storefronts that look visually similar through observing which particular environmental variables are atypical of the visual cluster this new image belongs to. A retail manger interested in siting a restaurant, for example, could photograph several prospective locations up for sale, and use our approach to retrieve a classification for each. The resulting classification would provide information describing whether the visual qualities of each location reflect typical uses of these spaces that are suited to their business. Following our restaurant example, a photograph classified as sharing visual commonalities to our Upmarket Hospitality cluster would likely present the most idealised location, by highlighting this photograph shares visual similarity to locations that appear to attract high volumes of hospitality services. In using this approach to complement existing tools, we argue taking into account the visual amenity of potential locations could help retail planners to arrive at *smarter* site location decisions, which carries wider implications for the vitality of town centres when amenities within these consumption spaces are optimally situated.

More generally, replication of our approach on a similar corpus of images (Google Street View, for example) could be used by planners to find whether different visual environments reflect particular patterns of built environment use, crime or socio-economic conditions of an area. Across particular urban centres, for example, planners might collect similar image-based data and apply our methods to identify visual commonality between different locations. Then, by collecting a set of variables of interest describing each location, planners might identify similarity or dissimilarity across different variables between the visual clusters. By example, if planners find a particular cluster suffers disproportionately high crime rates, they could sample a number of images from this cluster and undertake post-hoc analysis on possible visual cues embedded in images of these locations. In doing so, our approach provides means for planners to evaluate visual elements that potentially drive the incidence of conditions like crime, which might be identified from locations with high enclosure or no street lighting, for example.

A further contribution of the present study relates to several methodological innovations we introduce in the analysis. As our CAE model is unsupervised, it does not require large numbers of labelled images for training the model to produce visual features for each image. While the existing focus of the literature uses pre-trained or fine-tuned Convolutional Neural Networks (CNNs) for computer vision tasks in urban planning (Dubey et al., 2016; Seresinhe et al., 2017; Zhang et al., 2018), in the present paper we show that unsupervised techniques such as CAEs can also extract visual information from street-level imagery. This is advantageous for two reasons. Firstly, it does not require the user to assemble a large number of labelled images for training the CNN, which might possibly be derived from a non-expect workforce on a crowd-sourcing platform such as Amazon Mechanical Turk. And secondly, because pre-trained networks are often designed for a different purpose than that intended by the user, transfer learning approaches may provide sub-optimal performance if the images used are too heavily skewed compared to the data used to train the original network. Thus, while CNNs can be fine-tuned to the user's image data, a secondary contribution of this paper highlights the utility of CAEs for urban scientific tasks seeking to extract visual information from street-level imagery.

Despite these advantages, there exists conceptual and methodological limitations that frame the conditions for which the study should be interpreted. From a conceptual standpoint, it is reasonable to suggest the 15-minute walk catchment used to derive measures that describe the functional characteristics of the environment around each premise might not be reflective of reality on the ground. A 15-minute walk in a dense urban environment like London is likely to intersect a variety of scenes that possess polarised socio-economic and functional characteristics – for example, the short distance between the affluent and poorer areas of Clapham and Brixton, respectively. This means measures describing the built environment within each catchment might be

inaccurate due boundary effects that influence area consideration and create barriers beyond which consumers do not patronize. From a methodological perspective, a further limitation is that repeatability of the empirical approach is conditional on the availability of suitable GPU hardware for training the CAE model end-to-end. Unfortunately, deep learning models require appropriate hardware to train, and this presents a financial barrier of access to researchers interested in replicating (or extending) the empirical strategy to their own datasets. Despite these concerns, the main contribution of this article presents directions for future researchers to employ the deep learning methods adopted by the paper. As CAE networks are unsupervised, they offer flexibility to researchers seeking to extract visual features from image data without using pre-trained networks. This is a pertinent point to consider because the target domains of pre-trained networks are often purposed to answer a different research question than that asked by the user.

Credit authorship contribution statement

Sam Comber: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Daniel Arribas-Bel:** Supervision, Writing - review & editing, Funding acquisition. **Alex Singleton:** Supervision, Writing - review & editing, Funding acquisition. **Les Dolega:** Supervision, Writing - review & editing, Funding acquisition.

Appendix

A. Cluster validation

The lack of a single global optimization procedure is an inherent limitation of clustering exercises, meaning the plausibility and usefulness of the classification are typically split between the purpose it serves but also a validation of its system-wide accuracy. With this in mind, we pair human intuition for ascertaining a sensible number of clusters alongside a metric used for measuring cluster compactness known as average silhouette width. To determine the quality of possible cuts to the dendrogram and, therefore, resulting number of final clusters, we calculate the average silhouette width for several partitions of the 250-class k-means solution. Silhouette width ranges from $-1 \leq s_i \leq 1$, with higher values being desirable as they imply low within-cluster dissimilarity; it is calculated as $s_i = \frac{b_i - a_i}{max(a_i, b_i)}$, where a_i is the average Euclidean distance of i to all other data points in the same cluster, and b_i is the Euclidean distance of i to the cluster nearest to the one i is assigned to.

In practice, we average s_i for all observations for each cut from 2 to 249 of the dendrogram in Fig. A.1, taking the final cut as one that yields a high average silhouette and sensible number of clusters. By scanning the figure we are able to discern a sensible number of five clusters which is ideal because five is both manageable to describe and large enough unpack interesting between-cluster variation. To accompany this, we provide the resulting dendrogram for the five clusters in Fig. A.2, which visualises the agglomerative steps used to aggregate the 250-class k-means solution into five coarser groupings. This is important because hierarchical clustering techniques do not provide cluster partitions automatically, and so *tree-cutting* procedures are required to return partitions that reflect similarities amongst observations in the agglomerative procedure. In our case, while other cuts to the dendrogram offered reasonable performance, we take the decision to cut the dendrogram horizontally at this particular position (of the y-axis in Fig. A.2) because the five cluster solution has a high average silhouette width and sensible number of clusters.

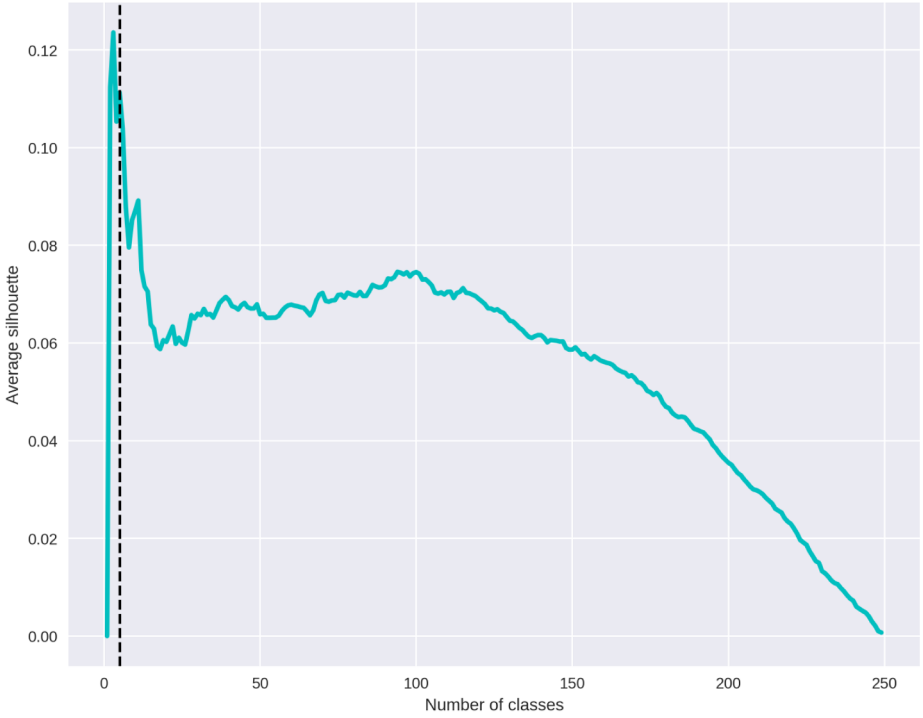

Fig. A1. Average silhouette for different aggregations of the 250-class k-means solution. Vertical dashed line indicates the desired five-class solution.

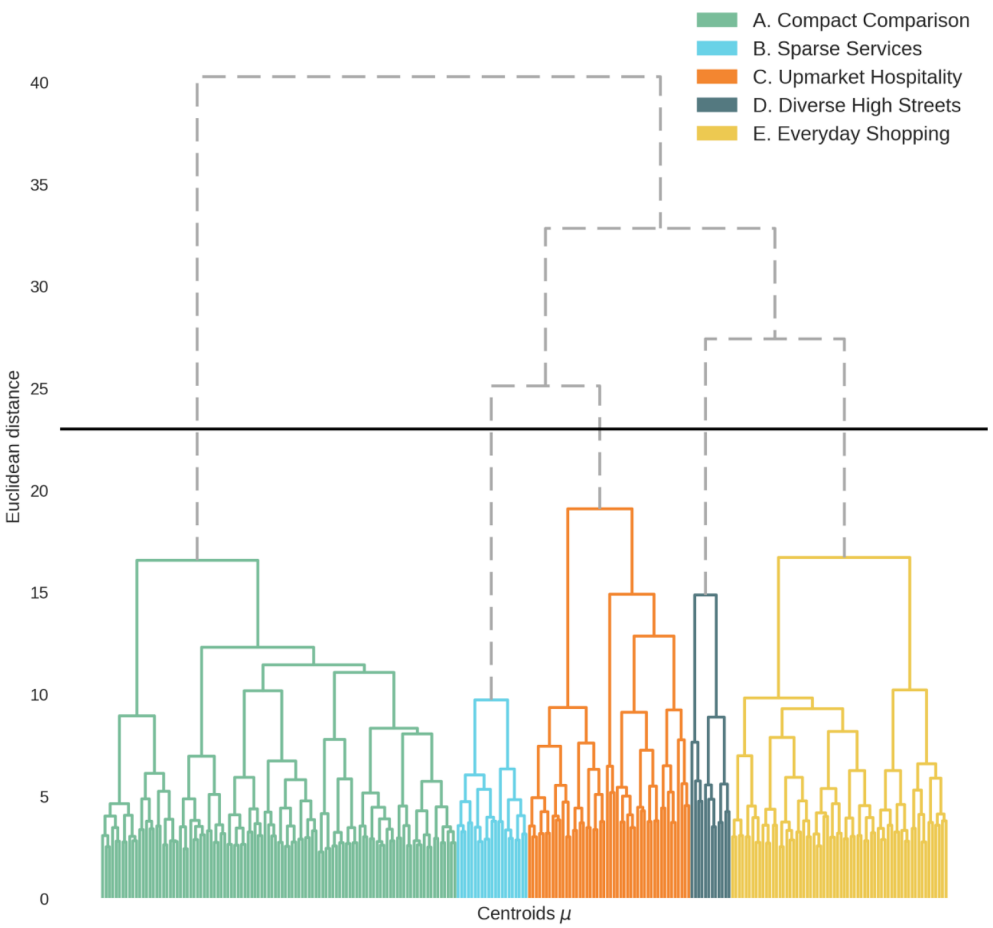

Fig. A2. Dendrogram displaying the agglomerative merge of the 250-class k-means solution.

B. Consistency with pre-trained visual features

To benchmark the visual features, h, retrieved from the latent representation encoded by the CAE we extract a similar set of visual features from two pre-trained Convolutional Neural Networks (CNNs): VGG16-Places365 (Kalliatakis, 2017) and ResNet50 (He et al., 2015). While pre-trained CNNs are trained using large volumes of labelled data for predicting a pre-defined set of categories, CAEs learn visual information that is optimised to the dataset supplied by the researcher. Between these approaches reflects a trade-off between the generalisability of CNNs to extract features learnt from a larger pool of images and more focused visual information extracted from the CAE trained on the researcher's data. Irrespective of this, both serve as points of comparison to assess the consistency of group memberships to particular clusters across different sets of visual features. Given these networks are pre-trained, they are not required to be trained from scratch, and so are initialized with existing weights. For VGG16-Places365, the network weights are initialized to those trained on the Places365 database consisting of 365 different environment categories – highways, vineyards, or libraries, for example – and are tuned for scene recognition tasks. ResNet50, on the other hand, is initialized with weights trained on the ImageNet database, which is a large visual dataset consisting of hand-annotated images that represent a wider range of 20,000 categories. For these pre-trained networks, we remove the fully-connected layer at the top of the network, meaning instead of returning probabilities for categories, we extract the visual features that are discriminative towards particular categories instead. In all, three sets of visual features are introduced to the clustering exercise introduced below. This includes visual features from the CAE represented by 784 pixels, VGG-Places365 features by 512 pixels, and ResNet50 features by 2048 pixels.

To externally validate our empirical approach we monitor changes in group membership and cluster sizes between visual features extracted from our CAE and the two pre-trained convolutional neural networks (CNNs), VGG16-Places365 and ResNet50. Thus, after clustering each set of visual features from the three models, we explore *agreeability* of cluster membership for a five cluster solution in Fig. B.1. The cluster sizes are represented by the vertical white rectangles for the CAE, VGG16-Places365, and ResNet50 models (left to right), with the frequency of leisure and retail amenities changing between groupings shown by the stream fields, and so represent changes in the composition of clusters between the three models. From an initial reading of the figure a mixed picture emerges. While the group sizes are moderately consistent between the CAE and VGG16-Places365, the clusters formed from the visual features of ResNet50 are far more balanced, with leisure and retail amenities spread more equally amongst the partitions. In regards to group membership, the highest agreeability is observable between the largest clusters partitioned using visual features of the CAE and VGG16-Places365 models. Similarly, the clusters identified by '0' in both models seem to share moderate agreeability, with there also being minor agreeability between '2' and '4' of the CAE and VGG16-Places365 models, respectively; the frequency flows of the remaining clusters are far more dispersed between different clustering solutions. Agreeability with ResNet50 visual features, on the other hand, is observably low, with there being no discernible patterns and consistencies between the clustering solutions. This is unsurprising given the target domain of both pre-trained networks is highly dissimilar, a phenomena known as data bias (Chen et al., 2017). While VGG16-Places365 is optimized for scene recognition tasks, ResNet50 is trained to predict over 20,000 object categories from the ImageNet database, with classes ranging from particular types of plants to bedroom items. The weights of the ResNet50 network are tuned to generate visual features that are discriminative for a wider range

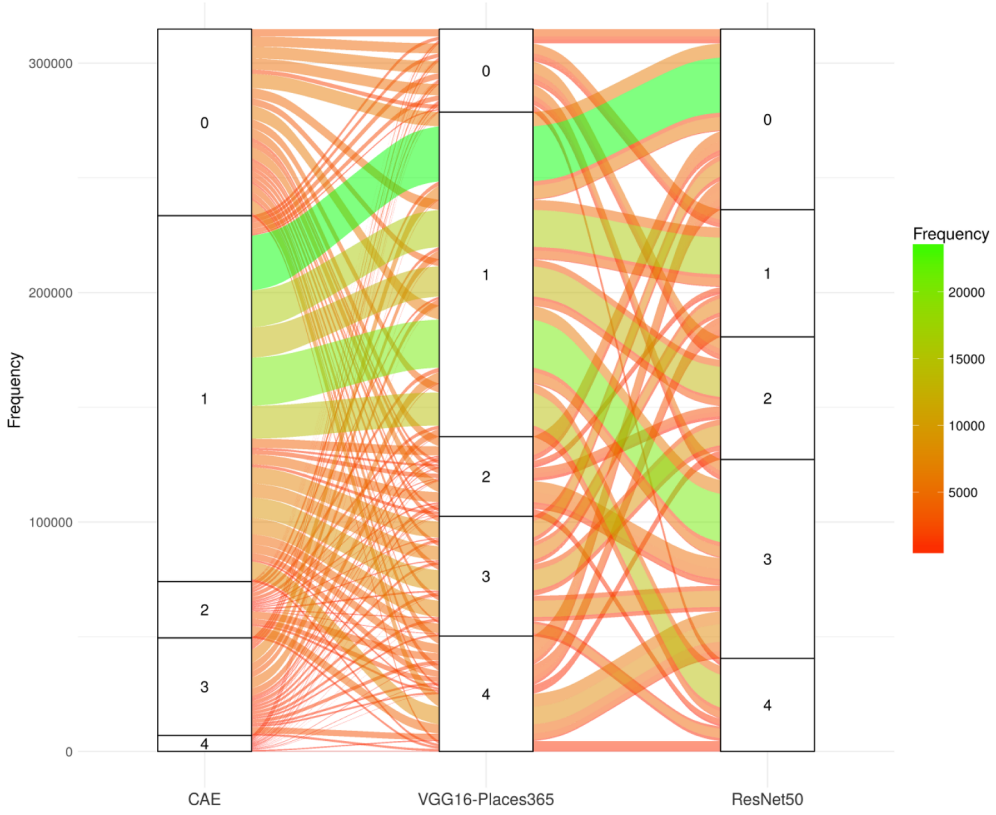

Fig. B1. Agreeability of the fiver cluster solution for visual features from Convolutional Autoencoder (CAE), VGG16-Places365, and ResNet50.

of object classes, meaning when we recover a representation for each leisure or retail amenity image, the kinds of features activated are more generalised than those from VGG16-Places365. This is due to the narrow focus for the range of categories that VGG16-Places365 has been trained to identify (with an emphasis on scene recognition tasks), meaning the visual features are more likely to be similar to those derived from the CAE model. Therefore, as the LDC images describe scenes observable from street-level, there is likely higher agreeability between the CAE and VGG16-Places365 models in terms of group membership and cluster sizes, which is reflected in the figure. All together, these observations confirm the visual features we extract using the CAE model are representing salient properties of the image, which motivates our descriptions for the characteristics of particular visual clusters in our empirical findings section.

References

Arribas-Bel, D. (2014). Accidental, open and everywhere: emerging data sources for the understanding of cities. *Applied Geography, 49*, 45–53.

Bader, M., Mooney, S., Bennett, B., & Rundle, A. (2017). The promise, practicalities, and perils of virtually auditing neighborhoods using google street view. *The Annals of the American Academy of Political and Social Science, 669*(1), 18–40.

Baker, J., Grewal, D., & Parasuraman, A. (1994). The influence of store environment on quality inferences and store image. *J Acad Mark Sci, 22*(September), 328–339.

Ballantine, P., Jack, R., & Parsons, A. (2010). Atmospheric cues and their effect on the hedonic retail experience. *International Journal of Retail & Distribution Management, 38*(June), 641–653. https://doi.org/10.1108/09590551011057453.

Bell, S. (1999). Image and consumer attraction to intraurban retail areas: An environmental psychology approach. *Journal of Retailing and Consumer Services, 6*(2), 67–78.

Bellizzi, J., Crowley, A., & Hasty, R. (1983). The effects of color in store design. *Journal of Retailing, 59*, 21–45.

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems, 65*, 126–139.

British Retail Consortium. (2009). "21st Century High Streets: A New Vision for Our Town Centres." http://www.boots-uk.com/media/3866/21st_century_high_streets_2012-1-1.pdf [Accessed 08/11/2019].

Carmona, M. (2015). London's local high streets: The problems, potential and complexities of mixed street corridors. *Progress in Planning, 100*, 1–84.

Chebat, J., Sirgy, J., & Grzeskowiak, S. (2010). How can shopping mall management best capture mall image? *Journal of Business Research, 63*(7), 735–740.

Chen, Y., Chen, W., Chen, Y., Tsai, B., Wang, Y., & Sun, M. (2017). No more discrimination: Cross city adaptation of road scene segmenters. *CoRR, abs/1704.08509*.

Chollet, François. (2015). "Keras." https://keras.io. [Accessed 08/11/2019].

Dennis, C., Marsland, D., & Cockett, T. (2002). Central place practice: Shopping centre attractiveness measures, hinterland boundaries and the Uk retail hierarchy. *Journal of Retailing and Consumer Services, 9*(4), 185–199.

Department for Transport. (2014). "National Public Transport Access Nodes (Naptan)." https://data.gov.uk/dataset/ff93ffc1-6656-47d8-9155-85ea0b8f2251/national-public-transport-access-nodes-naptan.

Dolega, L., Reynolds, J., Singleton, A., & Pavlis, M. (2019). Beyond retail: New ways of classifying Uk shopping and consumption spaces. *Environment and Planning B: Urban Analytics and City Science*.

Dubey, A., Naik, N., Parikh, D., Raskar, R., and Hidalgo, C. (2016). "Deep Learning the City : Quantifying Urban Perception at A Global Scale." CoRR abs/1608.01769. http://arxiv.org/abs/1608.01769.

Efron, B. and Hastie, T. (2016). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. 1st ed. New York, NY, USA: Cambridge University Press.

El-Adly, M. (2007). Shopping malls attractiveness: A segmentation approach. *International Journal of Retail and Distribution Management, 35*(11), 936–950.

Glaeser, E., & Gottlieb, J. (2009). The wealth of cities: Agglomeration economies and spatial equilibrium in the United States. *Journal of Economic Literature, 47*(4), 983–1028. https://doi.org/10.1257/jel.47.4.983.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge, MA, USA: MIT Press.

Hauser, J., & Koppelman, F. (1979). Alternative perceptual mapping techniques: relative accuracy and usefulness. *Journal of Marketing Research, 16*(4), 495–506.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Deep Residual Learning for Image Recognition." arXiv Preprint arXiv:1512.03385.

Inkscape Project. (2019). "Inkscape." https://inkscape.org. [Accessed 08/11/2019].

Jacobs, J. (1961). *The death and life of great american cities.* New York, NY, USA: Random house.

Jung, H., Lee, S., Kim, H., & Lee, J. (2017). Does improving the physical street environment create satisfactory and active streets? Evidence from Seoul's Design Street Project. *Transportation Research Part D: Transport and Environment, 50*, 269–279.

Kalliatakis, G. (2017). "Keras-Vgg16-Places365." https://github.com/GKalliatakis/Keras-

VGG16-places365; GitHub. [Accessed 08/11/2019].
Kant, I. (1790). Critique of Judgment. New York City, NY, USA: Barnes & Noble.
Kelling, G., & Coles, C. (1997). *Fixing broken windows: Restoring order and reducing crime in our communities.* A Touchstone Book: Free Press.
Liu, L., Zhou, B., Zhao, J., & Ryan, B. (2016). C-Image: City cognitive mapping through geo-tagged photos. *GeoJournal, 81*(6), 817–861.
Lynch, K. (1960). *The image of the city.* Cambridge, MA, USA: MIT Press.
Massey, D. (1991). A global sense of place. *Marxism Today, 38*, 24–29.
Merleau-Ponty, M. (2004). *The world of perception.* Abingdon, Oxford, UK: Routledge.
ONS. (2016) "National Records of Scotland; Northern Ireland Statistics and Research Agency (2016): 2011 Census Aggregate Data. UK Data Service (Edition: June 2016)." Office for National Statistics. https://census.ukdataservice.ac.uk/use-data/citing-data.aspx. [Accessed 08/11/2019].
-. (2019). "Introduction to Output Areas - the Building Block of Census Geography." 2019. https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas. [Accessed 08/11/2019].
Page, S., & Hardyman, R. (1996). Place marketing and town centre management: A new tool for urban revitalization. *Cities, 13*(3), 153–164.
Quercia, D., O'Hare, N., & Cramer, H. (2014). *Aesthetic capital: What makes london look beautiful, quiet, and happy?* (pp. 945–955). CSCW '14. New York, NY, USA: ACM. https://doi.org/10.1145/2531602.2531613.
Ravenscroft, N. (2000). The vitality and viability of town centres. *Urban Studies, 37*(13), 2533–2549.
Ryan-Collins, J., Cox, J., Potts, R., and Squires, P. (2010). "Re-Imagining the High Street - Escape from Clone Town Britain." New Economics Foundation. https://b.3cdn.net/nefoundation/1da089b4b1e66ba2b3_v8m6b0c0w.pdf. [Accessed 08/11/2019].
Salesses, K., Schechtner, K. and Hidalgo, C. (2013). "The Collaborative Image of the City: Mapping the Inequality of Urban Perception." PLOS ONE 8 (7). Public Library of Science:1–12.
Seresinhe, C., Preis, T., Moat, H. (2017). "Using Deep Learning to Quantify the Beauty of Outdoor Places." Royal Society Open Science 4 (7). The Royal Society. https://doi.org/10.1098/rsos.170170.
Silver, D., & Clark, T. (2016). *Scenescapes: How qualities of place shape social life.* Chicago USA: University of Chicago Press.
Singleton, A., Dolega, L., Riddlesden, D., & Longley, P. (2016). Measuring the spatial vulnerability of retail centres to online consumption through a framework of E-resilience. *Geoforum, 69*, 5–18.
Sorokin, A, Forsyth, D. (2008). "Utility Data Annotation with Amazon Mechanical Turk." In 2008 Ieee Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Cvpr Workshops. https://doi.org/10.1109/CVPRW.2008.4562953.
Spielman, S., & Singleton, A. (2015). Studying neighborhoods using uncertain data from the american community survey: A contextual approach. *Annals of the Association of American Geographers, 105*(5), 1003–1025.
Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.
Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning, 180*, 148–160.