ARTICLE

Measuring and modelling perceptions of the built environment for epidemiological research using crowd-sourcing and image-based deep learning models

Andrew Larkin[1], Ajay Krishna[2], Lizhong Chen[2], Ofer Amram[3], Ally R. Avery[3], Glen E. Duncan[3] and Perry Hystad[1]✉

BACKGROUND: Perceptions of the built environment, such as nature quality, beauty, relaxation, and safety, may be key factors linking the built environment to human health. However, few studies have examined these types of perceptions due to the difficulty in quantifying them objectively in large populations.

OBJECTIVE: To measure and predict perceptions of the built environment from street-view images using crowd-sourced methods and deep learning models for application in epidemiologic studies.

METHODS: We used the Amazon Mechanical-Turk crowdsourcing platform where participants compared two street-view images and quantified perceptions of nature quality, beauty, relaxation, and safety. We optimized street-view image sampling methods to improve the quality and resulting perception data specific to participants enrolled in the Washington State Twin Registry (WSTR) health study. We used a transfer learning approach to train deep learning models by leveraging existing image perception data from the PlacePulse 2.0 dataset, which includes 1.1 million image comparisons, and refining based on new WSTR perception data. Resulting models were applied to WSTR addresses to estimate exposures and evaluate associations with traditional built environment measures.

RESULTS: We collected over 36,000 image comparisons and calculated perception measures for each image. Our final deep learning models explained 77.6% of nature quality, 68.1% of beauty, 72.0% of relaxation, and 64.7% of safety in pairwise image comparisons. Applying transfer learning with the new perception labels specific to the WSTR yielded an average improvement of 3.8% for model performance. Perception measures were weakly to moderately correlated with traditional built environment exposures for WSTR participant addresses; for example, nature quality and NDVI ($r = 0.55$), neighborhood area deprivation ($r = -0.16$), and walkability ($r = -0.20$), respectively.

SIGNIFICANCE: We were able to measure and model perceptions of the built environment optimized for a specific health study. Future applications will examine associations between these exposure measures and mental health in the WSTR.

IMPACT STATEMENT: Built environments influence health through complex pathways. Perceptions of nature quality, beauty, relaxation and safety may be particularly import for understanding these linkages, but few studies to-date have examined these perceptions objectively for large populations. For quantitative research, an exposure measure must be reproducible, accurate, and precise––here we work to develop such measures for perceptions of the urban environment. We created crowd-sourced and image-based deep learning methods that were able to measure and model these perceptions. Future applications will apply these models to examine associations with mental health in the Washington State Twin Registry.

Keywords: Deep learning; perceptions; built environment; safety; nature quality; exposure assessment

Journal of Exposure Science & Environmental Epidemiology (2022) 32:892–899; https://doi.org/10.1038/s41370-022-00489-8

INTRODUCTION

There are many complex pathways through which the built environment may influence human health. These include modifying exposures to environmental hazards, such as air pollution [1] or heat [2], providing accessibility to essential resources, such as food and healthcare [3], modifying behaviors, such as physical activity [4], facilitating or mitigating pedestrian safety [5], and influencing stress [6] to name a few. Over the past two decades, numerous studies have observed associations between measures of the physical built environment and health outcomes [7]. For example, a systematic review of 36 longitudinal studies found that neighborhood walkability (assessed by land use mix, road connectivity, and destinations) was strongly associated with obesity, type 2 diabetes, and hypertension [8]. These studies have

[1]College of Public Health and Human Sciences, Oregon State University, Corvallis, OR, USA. [2]College of Engineering, Oregon State University, Corvallis, OR, USA. [3]Elson S. Floyd College of Medicine, Washington State University, Health Sciences Spokane, Spokane, WA, USA. ✉email: Perry.Hystad@oregonstate.edu

Received: 27 April 2022 Revised: 19 October 2022 Accepted: 20 October 2022
Published online: 11 November 2022

been essential for urban planners to maximize health benefits, and minimize health detriments, of built environments. However, our understanding of how the built environment influences health remains constrained by the methods currently available to measure urban characteristics objectively and at scales needed for population-based research.

Perceptions, defined here as a way of regarding and interpreting different built environments, may be a particularly important construct of the built environment that is challenging to assess with traditional measurement methods. Perceptions can include how safe individuals view an environment, the quality of green space, or the overall attractiveness or liveliness of an environment. These types of perceptions can influence diverse behaviors and mental health states, ranging from physical activity [9] to subjective well-being [10]. There are many psychological frameworks to support how perceptions of the built environment may influence health, such as the theory of planned behavior [11], social cognitive theory [12], and the social ecological model [13]. According to these theories, perception is a key factor to understanding the causal mechanism between the built environment, human behavior, and human health [14].

The widespread availability of street view images, combined with new prediction methods using deep learning models, offers an entirely new approach for measuring and modelling complex built environment constructs [15]. Street view images are images taken to capture the visible characteristics of a street. Many street view image datasets, such as Google Street View (GSV), Baidu and Tencent images are taken systematically to offer wide coverage across large geographic areas. Other street view datasets (e.g. Mapillary) are crowd-sourced, collected from voluntary participants. In addition to these available street-level imagery datasets, deep learning models have recently gained popularity–these models leverage millions of interconnected coefficients (weights) and non-linear equations (activation functions) to capture complex, non-linear associations. Together, image-based deep learning exposure assessments are emerging as a feasible approach for large-scale health studies [16]. Traditionally, time consuming and costly large-scale surveys, personal interviews, or field data collection are used to collect perceptions of the built environment, such as the quality of green space [17] or safety of neighborhoods [18]. This limits the application of these constructs to health studies. For example, most epidemiological studies examining the health benefits of urban green space use measures of the presence of green space (e.g., park locations, tree canopy cover, or the satellite-derived Normalized Difference Vegetation Index (NDVI)) [19], which do not capture the quality of green space or other associated complex constructs (e.g. safety, attractiveness) that could mediate the relationship between green space and health.

Recent research using web-based crowdsourcing methods and deep learning models show that complex perceptions can be measured and modelled [20]. Crowdsourcing methods have been used to collect perceptions via image comparison [21] or ranking of images by instruments (e.g. 1–10 scale) to quantify specific perceptions [22]. For example, the MIT Place Pulse (PlacePulse) study was a crowdsourced game for capturing image perceptions of the built environment [21]. Given a comparison of two images, participants chose which image is superior (i.e. wins) for a given perception of interest (e.g. safety, lively). The PlacePulse study collected more than 1.1 million votes from participants across the globe and results showed that online images can be used to create reproducible measures of urban perception and characterize differences between cities [21]. In addition, deep learning models built upon the PlacePulse data were able to explain a large component of these perceptions [21], although more data on local, within-city perceptions are needed to capture localized built environment perceptions most important to health [23].

For quantitative research, an exposure measure must be reproducible, accurate, and precise––here we work to develop such a measure for perceptions of the urban environment. We developed methodology to measure and predict perceptions of the built environment using street-view imagery, crowd-sourced image comparisons, and deep learning models. We focused on quantifying built environment perceptions of nature quality, beauty, relaxation, and safety for application in an epidemiologic cohort study (the Washington State Twin Registry (WSTR)) [24] to examine how urban green space is associated with mental health. We chose to examine these four perceptions based on how we conceptualize green space may influence mental health [25], and based on existing perception data in the PlacePulse dataset [21]. We describe methodological approaches to optimize perception data collection when using crowdsourced methods specific for an epidemiological study and develop transfer learning deep learning models to predict nature quality, beauty, relaxation, and safety for street-view images. We applied these models to residential locations of WSTR participants to determine if our perception measures are distinct from traditional built environment exposure measures used in health studies.

METHODS

Google street-view image training dataset

We focused on collecting perception data for GSV images in Washington State, which contains nearly 80% of addresses collected by the WSTR. A total of 117,374 GSV images were collected from the GSV API. Images were restricted to the 2010 US Census Urbanized Areas (UA) and Urbanized Clusters (UC) (Supplemental Figs. 1, 2). Images were sampled from each UA and UC proportional to the percentage of WSTR participants residing within each UA or UC. GSV image viewing angles were randomly chosen at 90° increments, and classified as straight (0° and 180°) or side (90° and 270° relative to vehicle heading).

We used the deep learning model PSPNet to quantify features in each image. PSPNet classifies each pixel of an image as one of 140 anthropogenic (e.g., road, building) or natural (e.g., tree, sky) objects [26]. We also grouped pixel classifications into themes including accessibility, green space, blue space, built environment, housing, and animate (transient objects). Themes and their corresponding set definitions are listed in Supplemental Table 1.

Street view images were grouped into subsets based on viewing angle and urban area classification. Seattle was considered a separate group from other UAs due to its high percentage of WSTR residents (50%). Within each group, we selected a subset of images (~1%) which maximized the uniform distribution of environmental composition calculated from the PSPNet algorithm (Supplemental Figs. 3–8).

$$Selected\ Subset = argmin_a \sum\nolimits_{j=1}^{n} \sum\nolimits_{k=1}^{10} \left| O_{jk} - E_{jk} \right| \qquad (1)$$

Where

a is the subset of images within partition A

j is the jth theme or object

n is the number of themes and objects considered in the statistic

k is the range of percent pixels from (k–1)*10 to k*10–1. For example, for $k = 1$ the range of percent pixels is 0–9%.

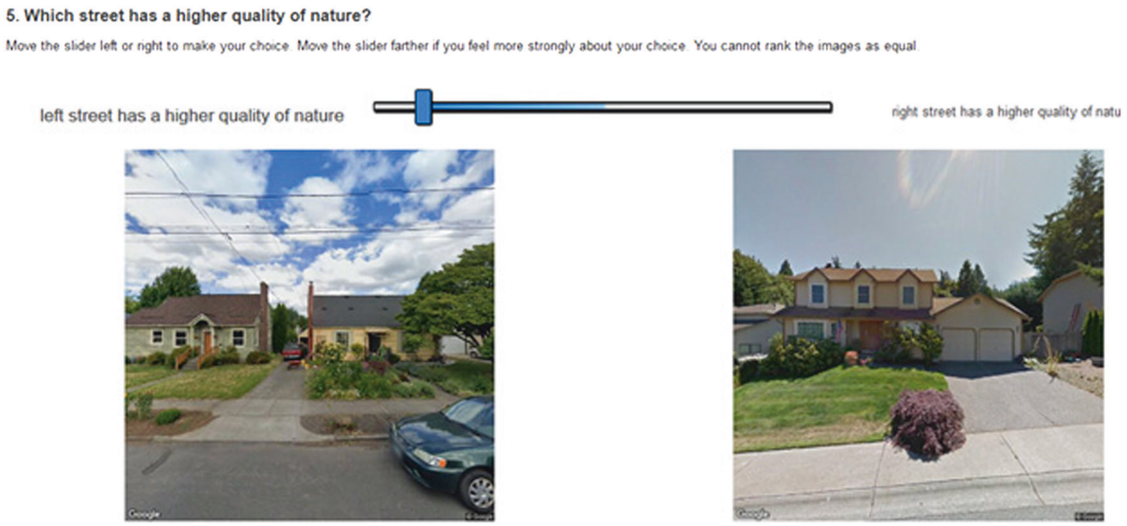
O_{jk} is the observed number of images within subset a whose percent pixel count for subset j falls within range k

E_{jk} is the expected number of images, given a uniform distribution, within subset a whose percent pixel count for subject j falls within range k

Collecting perception data

We collected perception data through Amazon Mechanical Turk (AMT). AMT is a service which pays participants, called 'Turkers', a small amount of money for completing short tasks, called 'HITS' (https://www.mturk.com/). Metadata for each HIT includes the time accepted, time completed, Turker id and state (based on IP address). Participants were paid on average $0.20 for each survey.

Surveys consisted of five image comparisons, one for each perception in our study (nature quality, beauty, relaxation, and safety), and a quality control question where participants chose which of two images had more cars (Supplemental Fig. 9). The order of perceptions and position of images on the left or right of image comparisons was randomly generated for each survey. Rather than clicking on the chosen image as in the preceding



5. Which street has a higher quality of nature?

Move the slider left or right to make your choice. Move the slider farther if you feel more strongly about your choice. You cannot rank the images as equal.

left street has a higher quality of nature
right street has a higher quality of natu
Google
Google

Fig. 1 Example of the Mechanical Turk platform where perception data was collected. Here participants are shown two images and asked "which street has a higher quality of nature?". Participants used the slider bar to select what image they thought had higher nature quality and how strongly they felt about this choice.

PlacePulse study [21], participants used a slider to convey their selection and moved the slider farther left or right based on how strongly they felt about their choice (Fig. 1). To remove potential biases with comparing images taken from different angles, all images within each survey contained the same viewing angles (straight or side). Surveys were valid if participants correctly answered a QA question (which image has more cars). Surveys completed in less than 30 s had significantly greater rates of failing the QA question ($p < 0.001$). We therefore excluded surveys completed in less than 30 s. Participant inclusion criteria included 18 years of age or older and living within the US

Calculating perception scores

The TrueSkill algorithm [27] used in the PlacePulse study was created by Microsoft to predict outcomes of video game competitions between two opposing players. Each player is represented by two variables: predicted skill level (mu) and uncertainty in predicted skill level (sigma). New players start with (mu, sigma) values of (25,6). Given (mu, sigma) values for two competing players, TrueSkill predicts which player will win. As players compete against each other, mu scores increase or decrease for wins and losses, respectively, and sigma values monotonically decrease. In PlacePulse, each image comparison is a competition between two images (players) for which image has a greater value for a specific perception (e.g. 'beauty'). Given an infinite number of comparisons, image mu scores will eventually settle around their true perception value.

We created a multinomial version of the TrueSkill algorithm, which also considers the magnitude of how much the chosen image in a comparison is preferred by the participant. Each survey the participant completes is transformed into three simulated games for slight, moderate, or strong intensity wins. When a participant chooses an intensity less than the minimum threshold for a greater intensity game, then the greater intensity game is recorded as a tie between the comparison images. The overall TrueSkill score and uncertainty is the sigma weighted average of the three simulated TrueSkill scores and the arithmetic mean of the simulated TrueSkill uncertainty, respectively.

$$\sigma_x = \frac{\sigma_{slight} + \sigma_{modereate} + \sigma_{strong}}{3} \tag{2}$$

$$TS_{avg} = \frac{\sigma_{slight} * TS_{slight} + \sigma_{moderate} * TS_{moderate} + \sigma_{strong} * TS_{strong}}{3\sigma_x} \tag{3}$$

Where

σ_x = average TrueSkill uncertainty

$\mathrm{TS}_{\mathrm{avg}}$ = multinomial TrueSkill score

In the PlacePulse study, deep learning model performance was dependent on the similarity of images being compared [21]. For comparisons in which images had similar TrueSkill scores, model performance did not differ from random guessing. To properly train a sensitive deep learning perception model, the training dataset requires a disproportionately greater number of comparisons between similar rather than differing images. We developed an algorithm that sampled image comparisons in proportion to the similarity of image TrueSkill scores (Supplemental Methods 1, Supplemental Eqs. 1, 2, Supplemental Fig. 11). We updated image TrueSkill scores in real time as surveys were submitted and sampled image comparisons in real time as surveys were requested by AMT participants.

Participants had varying individual tendencies for image comparison magnitudes. To adjust for personal tendencies, we standardized magnitudes using a weighted average of all image comparisons within a survey (Supplemental Methods 2, Supplemental Eqs. 3–5, Supplemental Figs. 12, 13).

Developing transfer learning deep learning models

The PlacePulse deep learning perception model architecture is based on two parallel Convolutional Neural Networks (CNNs) that extract features from each image and compares features to choose the winning image (Supplemental Fig. 14). We simplified the PlacePulse architecture to predict a continuous rather than a binary score. The feature extractor layers were retrained, the ranking subnetworks were removed, and the binary SoftMax layer in the fusion subnetwork output was replaced with a continuous output (mean squared error loss) (Fig. 2).

We leveraged the original PlacePulse image comparison dataset to conduct transfer learning with our new image labels specific to the WSTR. Transfer learning is a deep learning approach in which model coefficients are first fitted to a large dataset related to the topic of interest and then fine-tuned using a smaller but more relevant dataset. We pretrained our perception models using PlacePulse image comparisons, replacing the final layer in the fusion subnetwork before fine-turning model weights with our survey records. Our safe model was pretrained with PlacePulse comparisons for 'safe', while our beauty, relaxing, and greenspace quality models were pretrained with PlacePulse comparisons for 'beautiful'. Sensitivity analyses include stratifying model performance by image viewing angle, urban classification, voting intensity, and winning image position (left or right).

Applying and evaluating measures in the WSTR Study

We collected GSV images for 31,607 WSTR participant addresses, capturing the closest street-view image to each address and survey date as well as five additional images within 100 meters (to assess neighborhood street-view measures). For comparison, and to determine if our GSV perception measures are distinct from traditional built environment exposure measures used in health studies, we calculated a number of built environment exposure variables. These included: summer (June and September) NDVI within 100 m of residential addresses using Landsat 30 m images; a walkability score derived from the EPA's smart location database, a nationwide geographic data resource for measuring location efficiency for the year 2013 [28]; an urban sprawl index derived [29]; population density within 1 km of residences

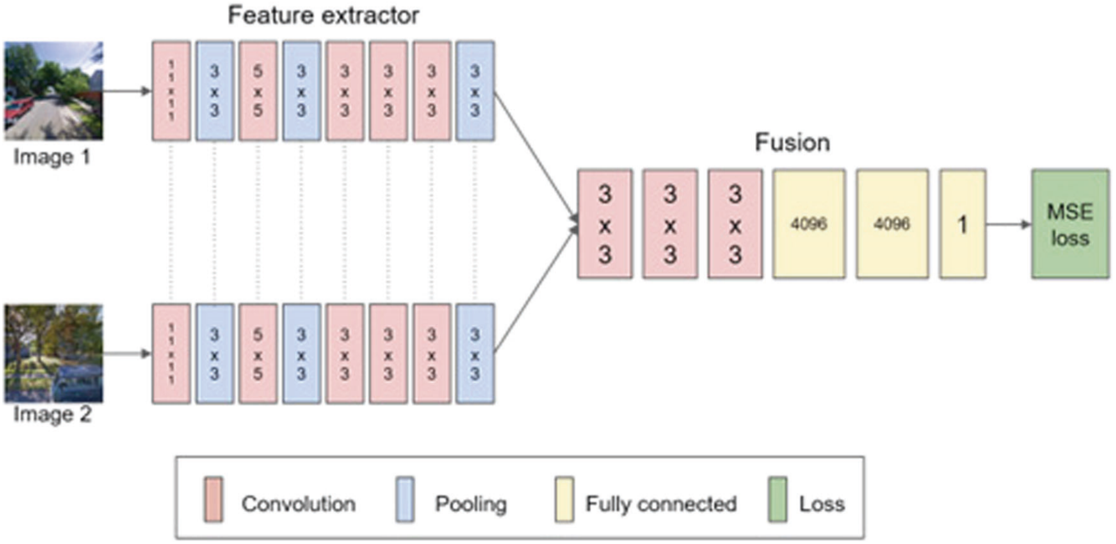


Fig. 2 Architecture of the deep learning model used to predict continuous perception scores for GSV images.

Table 1. Mean and standard deviation of select TrueSkill performance metrics.

	Score[a] (sd)	Intensity (sd)	Uncertainty (sd)
Nature quality	24.95 (4.24)	3.49 (2.41)	2.53 (0.32)
Safety	24.99 (2.86)	2.29 (1.72)	2.38 (0.31)
Beauty	25.03 (3.50)	2.83 (2.05)	2.45 (0.32)
Relaxation	25.03 (3.77)	3.03 (2.23)	2.49 (0.33)

[a] Multinomial TrueSkill score. Intensity is the average magnitude of participant votes. Uncertainty is analogous to the binary TrueSkill parameter mu.

derived from 2015 census block points; and an Area Deprivation Index (ADI) [30]. We also explored associations with the PSPNet calculations of % green space, % trees, % blue space, and % built up area within the GSV images. We tested for spatial autocorrelation by calculating Moran's I with the Spatial Autocorrelation function in ArcPro (2.9.3) using euclidean inverse distance weighting to define spatial relationships (default settings).

RESULTS

Perception measures

A total of 9,192 surveys were collected from Mechanical Turk. After applying our exclusion criteria (QA/QC question and time spent on survey) there were 8085 surveys from 4932 unique participants. The estimated error rate in the screened dataset is 1.5% (Supplemental Fig. 15, Supplemental Methods 3). More than 75% of comparisons ($n = 6074$) were from participants who completed three or fewer surveys (Supplemental Table 2). The number of completed surveys by state was proportional to US population density (Supplemental Fig. 16, Supplemental Table 3).

Calculated TrueSkill metrics for nature quality, safety, beauty and relaxation are summarized in Table 1. Morans' I spatial clustering values for TrueSkill scores and bias are shown in Fig. 3, stratified by urban category. Clustering ranges from 0.94 (relaxing in Seattle) to −0.20 (beauty bias in UC neighborhoods). Clustering is greater for TrueSkill scores across Seattle (0.34–0.84) and UAs (0.44–0.88) compared to UCs (0.09–0.26). However, weaker clustering may be due to smaller UC sample size and fewer nearby neighbor comparisons within the 1 km Moran's I threshold. Nature quality and relaxing perceptions are more strongly clustered within Seattle and UA neighborhoods than safety and beauty. Spatial clustering is weaker for bias (range −0.2 to 0.27).

Deep learning model performance

Performance metrics of the continuous perception models and the original PlacePulse models in a validation dataset are shown in Table 2. Metrics are based on binary evaluation metrics (did the model correctly identify the winning image) since PlacePulse outputs are binary. For both the binary and continuous models, performance was greatest for nature quality (72.9–77.6%) and lowest for safety (63.5–64.4%). Comparing the continuous and binary PlacePulse models, performance metrics are consistently greater for the continuous models, with metric improvements ranging from 0.9% (safety) to 5.6% (relaxing).

Model performance stratified by image position, viewing angle, urban development level, and vote intensity is shown in Table 3. While model performance is not affected by image position (left or right), images with a straight viewing angle have 4% greater performance for relaxation, beautiful, and nature quality perceptions compared to a side viewing angle. Model performance is 7.3% (nature quality) to 12.5% (beautiful) higher for Seattle and other UAs compared to UCs. Model performance also significantly varies by voting intensity. Performance for comparisons between similar images (1–10 intensity) is 29.4% (safety) to 54.1% (beautiful) lower than comparisons between dissimilar images (11–50 intensity). The impact on the modelled TrueSkill score between unadjusted and adjusted perception votes was minimal (Supplemental Fig. 17).

Application to the WSTR

A total of 24,354 out of 31,607 addressees (77%) had GSV images within 100 m of home; 25,730 twin addresses were located in urban areas, of which 21,566 (84%) had GSV images within 100 m. We explored locations with missing images, which resulted for

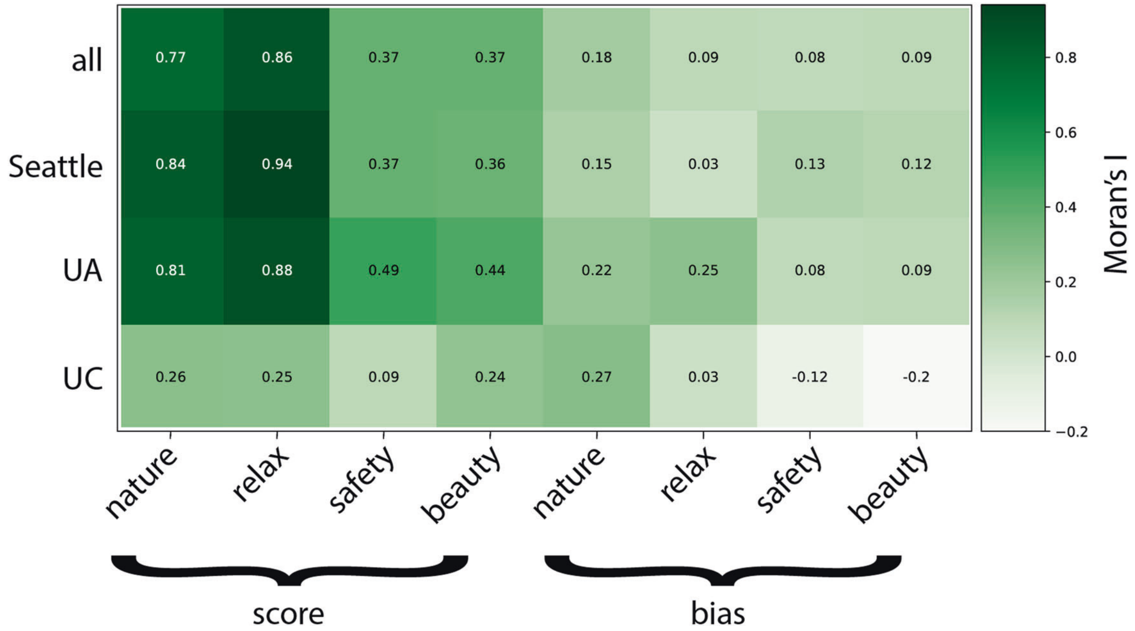


Fig. 3 Morans' I values for adjusted TrueSkill scores and bias, for the entire dataset and stratified by urban size category (Seattle images, urban area (UA) images, and urban clusters (UC) images).

Table 2. Performance of the updated continuous models compared to the original place pulse models.

Perception	Place pulse (Binary)	Continuous models
Nature quality	72.9%	77.6%
Safety	63.5%	64.4%
Relaxing	66.4%	72.0%
Beautiful	63.7%	68.1%

Table 3. Stratification of perception training dataset and model performance.

	Nature Quality	Safety	Relaxation	Beauty
Winning image position				
Left	77.9%	63.9%	72.3%	68.4%
Right	77.4%	64.9%	71.6%	67.5%
Street Image Viewing Angle				
Straight	79.5%	63.1%	75.0%	70.0%
Side	75.5%	65.7%	69.0%	66.0%
Urban Level				
Seattle	81.1%	63.0%	72.4%	67.8%
Urban Area (>50,000 people)	77.8%	65.7%	75.9%	69.7%
Urban Cluster (2,500–49,999 people)	70.5%	53.3%	64.1%	55.3%
Vote intensity				
1–10 votes	15.1%	18.8%	14.7%	6.8%
11–50 votes	68.2%	48.2%	57.1%	60.9%

many reasons, such as gated communities, newly built communities, or inaccessible roads. The mean (std) predicted TrueSkill scores for study participants addresses was 25.05 (3.34), 25.11 (2.25), 25.12 (2.58) and 25.19 (2.01) for nature quality, beauty, relaxation, and safety, respectively. The correlation between perceptions ranged from a low of 0.15 (safety and beauty) to a high of 0.70 (relaxation and nature quality) (Table 4). The correlations between the nearest, average, and IDW of all images within a 100 m buffer of a home were all >0.77, suggesting similarity in GSV measures for neighborhood areas around homes. We therefore used the nearest images to the residence for comparisons between perceptions, image features, and other built environment exposures

The correlation between perceptions, image features, and other built environment exposures is shown in Table 4. Image features, especially the percent of the image that was green space or trees contributed to nature and beauty perceptions, while built up features were positively associated with safety. The correlation between our predicted perceptions and satellite NDVI (within 100 m) were low (−0.04 for safety) to moderate (0.55 for nature quality) (Table 4). Increasing levels of walkability, sprawl, population density, and area deprivation were all associated with decreasing perceptions, with correlation coefficients below 0.38.

DISCUSSION

Measures of the built environment tend to focus on physical characteristics, such as road density, tree cover, or temperature. Perceptions of the built environment, such as urban nature quality, beauty, relaxation, or safety, are unique constructs that may be particularly important for understanding complex linkages between the built environment, human behavior, and human health. Here, we developed methods that successfully measured and modelled perceptions of the built environment optimized for the WSTR health study.

Measuring perceptions of different built environments for large geographical areas is challenging. Crowd sourced data collection designed with standardized quantitative questions has significant potential for advancing measurement of perceptions at the population level. We found that many decisions regarding survey design and sampling strategies significantly influenced the quality of the crowd sourced data we collected. Specifically, dynamically updating image sampling weights to ensure comparison of similar types of built environments ensured that we were not just capturing

Table 4. Correlation between predicted perceptions and measures of environmental exposures in the Washington State Twin Registry.

	Perceptions				Image composition				Built environmental measures				
	Nature quality	Beauty	Relax-ation	Safety	Green space (%)	Trees (%)	Blue space (%)	Built Up (%)	NDVI	Walk-ability	Sprawl index	Pop. density	ADI
Nature quality	1												
Beauty	0.58	1											
Relaxing	0.70	0.56	1										
Safety	−0.26	0.15	0.05	1									
Green space (%)	0.57	0.53	0.33	−0.06	1								
Trees (%)	0.60	0.57	0.28	−0.19	0.79	1							
Blue space (%)	0.05	0.03	0.04	−0.06	0.09	0.08	1						
Built Up (%)	−0.61	−0.24	−0.59	0.35	−0.02	−0.11	−0.05	1					
NDVI	0.55	0.47	0.44	−0.04	0.44	0.51	0.04	−0.36	1				
Walkability	−0.20	−0.24	−0.38	−0.12	−0.11	−0.07	−0.03	0.25	−0.44	1			
Sprawl index	−0.19	−0.23	−0.35	−0.13	−0.12	−0.1	−0.01	0.2	−0.47	0.79	1		
Pop. density	−0.17	−0.17	−0.3	−0.15	−0.11	−0.08	−0.02	0.25	−0.35	0.49	0.54	1	
ADI	−0.16	−0.16	−0.07	0.05	−0.1	−0.16	−0.03	0.04	−0.22	0.06	0.08	−0.04	1

extreme differences in perceptions, but that our measures (and resulting models) were able to distinguish between common built environments experienced everyday among WSTR participants. This was a major limitation of the original PlacePulse study, which was primarily driven by between city and region differences [23]. In addition, GSV images were selected so the entire dataset followed uniform distributions for important physical features of the built environment (e.g., percent of the images that were trees, buildings, roads, etc.). Adjusting survey responses based on individual voting tendencies also proved valuable in generalizing perceptions across a diverse population. For future studies, we propose splitting safety into two constructs: pedestrian safety and safety from crime. Defining, identifying, and measuring inter-individual differences is an important component of population level perception assessment that needs further development. Collecting demographic information from participants may help identify groups and cultures that tend to diverge or disagree from the general population, which may require different deep learning models to predict perceptions. Future research should develop best practices to ensure consistency and compatibility across crowdsourced studies and datasets.

The goal of our research was to develop models that could predict perceptions of the built environment for any study participant address within the WSTR. Deep learning models based on CNNs are ideal for capturing the complex, non-linear spatial relationships that are needed to predict perceptions from images. For image comparisons, CNN models such as the SIAMESE network identify and amplify important spatial differences within comparisons that are associated with our perception labels. The deep learning models we developed are significant improvements over the previous PlacePulse models [21]. By making key changes to the architecture of the original model we can now capture not only which image participants prefer, but also the magnitude of their preference. Recently, there have been several new deep learning models published to predict environmental features or perceptions from street-view images. For example, a CNN to predict whether an image is 'scenic' or not [31], replicating the PlacePulse model in Chinese Cities using Tencent Street View [32], and developing quantitative measures of cleanliness, comfort and traffic with a CNN model trained from Baidu Street View images [33]. Similar to developing best practices for dataset collection, developing best practices for model evaluation is important for comparing and contrasting deep learning perception models. Standardized test datasets for popular perceptions, such as beauty, should be collected with representation from diverse populations and central repositories developed where popular models can be curated, independently evaluated, and publicly available.

We optimized our training data to be able to predict perceptions for participants in the WSTR. A total of 24,354 addresses (77%) had GSV images within 100 m, with more missing data in rural compared to urban locations. There is a growing literature that has applied GSV images to assessing environmental exposures in health studies [34, 35], but these are typically measuring objects within images (e.g., cars, trees, people) rather than complex perceptions like nature quality, safety, or beauty. Here we assessed whether our new modeled perceptions of green space quality, beauty, relaxation, and safety were in fact distinct qualities of the built environment not captured by objects within GSV images (e.g., % cars, trees, people) and traditional built environment measures (NDVI, walkability, sprawl, population density, and area deprivation). Overall, we observed moderate to low correlations between these metrics. For example, our measure of nature quality was moderately correlated ($r = 0.55$) with the commonly used NDVI measure to assess green space exposure, as well as the percent of the GSV image that was green space ($r = 0.57$) or trees ($r = 0.60$). This suggests that our perception measures are in fact reflecting complex constructs of the built environment not captured by simple measures of presence or amount of green space features. Future research will explore how to combine different metrics (e.g., presence of green space + quality of nature) to best represent exposures most important to health.

Perceptions of the built environment, compared to objective measures of single (or single category) built environment features, presents new opportunities for environmental epidemiologists to examine complex relationships between cumulative built environment elements and their impacts on health. Research shows that multiple built environment exposures are inter-related and that the entirety of these environmental exposures are likely important to health [36–39]. This concept has been referred to as the exposome––a complement to the genome that captures the totality of environmental exposures from conception throughout the lifecourse [40, 41]. The objective perception measures we developed here can be viewed as one approach to measuring an overall built environment construct that could be important to different health-related behaviors, exposures, and health outcomes. Creating exposure measures based on individual built environment components and linear relationships will not capture these complex perceptions, as we have demonstrated in a previous study [23]. Deep learning models of imagery are needed to predict these perceptions. Importantly, the widespread availability of street-view imagery means that these models can be applied to residential addresses in diverse epidemiological studies, providing new, cumulative measures of built environment exposures. The type of exposure measures that can be created from images and deep learning models are endless [16], but these should be driven by the specific research question being examined and knowledge about the exposure pathways of interest. For example, our overall study goal is to examine how urban green space influences mental health in the WSTR, hence our focus on green space quality and streetscape safety, attractiveness and relaxation––important missing measures in the current green space epidemiological literature [25].

LIMITATIONS

Several limitations to our measurement and modelling methods should be considered with future applications. First, information about a place captured by a street-view image is limited to only streetscapes and does not include other built environment areas or perceptions related to sound or smell. There are also slight variations in GSV image quality (i.e., contrast, 57 hue, saturation, brightness, tint, and clarity), as well as weather conditions, season, and time of day that can introduce additional variation in the perceptions associated with an image. Second, while online survey methods and the AMT platform allowed us to reach participants across the US, we had trouble recruiting participants within the study area of interest (Washington State). Future improvements include expanding the developed methodology to alternative platforms (e.g., Survey Monkey) or offering greater rewards to incentivize participation in targeted regions. Longer sampling windows are also likely to recruit more participants. The trade-off between participant representation, accuracy, and sample size is also important to consider in future perception data collection. Third, in our collected surveys we adjusted for, rather than integrating, inter-individual differences in perception magnitudes. Traditional Likert-scale surveys often include demographic variables such as age, gender, income, and education, which are associated with differences in perception. Fourth, we collected participant perceptions from April 17th to May 13th, 2020. During this time, participants across the US were under stay-at-home orders to reduce the spread of COVID-19. It's uncertain how much the collected data accurately portrayed the past or can predict future perceptions. It is possible participant perceptions are driven by an anchor effect, where preference for one image over another is intuitive and consistent over time, and the magnitude of the preference is dynamic. Under this hypothesis, collected perceptions would still represent long-term differences in perception, as TrueSkill scores are relative rather than absolute measures. Fifth, the estimated error rate before and after implementing QA methods was 5.9% and 1.5%, respectively. Model sensitivity to error rates are dependent on the predictive power of the model. For example, in a training dataset with 5.9% sample error, the best accuracy any model can hope for is 94.1%. Additional accuracy would be overfitting. In these early stages of perception modeling, 5.9% error is not a concern; however, as models become more refined, high single digit error rates will likely have a negative impact on model performance. Sixth, GSV images were taken at different times throughout the year, most in summer, with repeated measures at most every year (with much lower temporal updates in rural and lower density areas). GSV exposure measures are therefore reflective of these time-periods. However, in previous research we observed that perceptions were relatively stable when applied to different GSV images over an approximate ten-year period, suggesting our new perception measures likely capture chronic-related, compared to acute, exposures [23]. Finally, GSV images within 100 m were available for 77% of all study residences (84% in urban areas). While no statistically significance differences ($p < 0.05$) were observed between sociodemographic characteristics for individuals with and without images, careful assessment of missingness is required for subsequent health analyses. Image missingness could also be reduced by increasing the distance from residential locations to the nearest GSV image (e.g. 250 m).

CONCLUSIONS

Perceptions are important pathways through which the built environment may influence behavior, psychological states, and ultimately health and wellbeing. We created methods to measure and model perceptions of the built environment that are optimized for epidemiological research. Future applications will apply these models to examine associations with mental health in the WSTR.

DATA AVAILABILITY

All scripts, database operations, and statistical analyses described in the methods section are available at the GitHub repository (https://github.com/larkinandy/Perceptions_MTurk). The WSTR dataset is not publicly available due to individual privacy concerns, but researchers can apply to access the WSTR data at https://wstwinregistry.org/.

REFERENCES

1. Voordeckers D, Lauriks T, Denys S, Billen P, Tytgat T, Van, et al. Guidelines for passive control of traffic-related air pollution in street canyons: an overview for urban planning. Landsc Urban Plan. 2021;207:103980.
2. Leal Filho W, Echevarria Icaza L, Emanche VO, Quasem Al-Amin A. An evidence-based review of impacts, strategies and tools to mitigate urban heat islands. Int J Environ Res Public Health. 2017;14:1600.
3. Gelormino E, Melis G, Marietta C, Costa G. From built environment to health inequalities: an explanatory framework based on evidence. Preventive Med Rep. 2015;2:737–45.
4. Kärmeniemi M, Lankila T, Ikäheimo T, Koivumaa-Honkanen H, Korpelainen R. The built environment as a determinant of physical activity: a systematic review of longitudinal studies and natural experiments. Ann Behav Med. 2018;52:239–51.
5. Stoker P, Garfinkel-Castro A, Khayesi M, Odero W, Mwangi MN, Peden M, et al. Pedestrian safety and the built environment: a review of the risk factors. J Plan Lit. 2015;30:377–92.
6. Beemer CJ, Stearns-Yoder KA, Schuldt SJ, Lowry CA, Postolache TT, Brenner LA, et al. A brief review on the mental health for select elements of the built environment. Indoor Built Environ. 2019. https://doi.org/10.1177/1420326X19889653
7. Hu M. Factors That Impact Human Health in the Built Environment. In: Hu M (ed). *Smart Technologies and Design For Healthy Built Environments.* Springer International Publishing: Cham, 2021, pp 37–56.

8. Chandrabose M, Rachele JN, Gunn L, Kavanagh A, Owen N, Turrell G, et al. Built environment and cardio-metabolic health: systematic review and meta-analysis of longitudinal studies. Obes Rev. 2019;20:41–54.
9. McGinn AP, Evenson KR, Herring AH, Huston SL, Rodriguez DA. Exploring associations between physical activity and perceived and objective measures of the built environment. J Urban Health. 2007;84:162–84.
10. Kent JL, Ma L, Mulley C. The objective and perceived built environment: What matters for happiness? Cities Health. 2017;1:59–71.
11. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process. 1991; 50:179–211.
12. Bandura A. *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall, Inc: Englewood Cliffs, NJ, US, 1986.
13. Sallis JF, Floyd MF, Rodríguez DA, Saelens BE. Role of built environments in physical activity, obesity, and cardiovascular disease. Circulation. 2012;125:729–37.
14. Ma L, Cao J. How perceptions mediate the effects of the built environment on travel behavior? Transportation. 2019;46:175–97.
15. Biljecki F, Ito K. Street view imagery in urban analytics and GIS: a review. Landsc Urban Plan. 2021;215:104217.
16. Weichenthal S, Hatzopoulou M, Brauer M. A picture tells a thousand…exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. Environ Int. 2019;122:3–10.
17. Knobel P, Dadvand P, Alonso L, Costa L, Español M, Maneja R. Development of the urban green space quality assessment tool (RECITAL). Urban Forestry Urban Green. 2021;57:126895.
18. Furr-Holden CDM, Campbell KDM, Milam AJ, Smart MJ, Ialongo NA, Leaf PJ. Metric properties of the neighborhood inventory for environmental typology (nifety): an environmental assessment tool for measuring indicators of violence, alcohol, tobacco, and other drug exposures. Eval Rev. 2010;34:159–84.
19. Rojas-Rueda D, Nieuwenhuijsen MJ, Gascon M, Perez-Leon D, Mudu P. Green spaces and mortality: a systematic review and meta-analysis of cohort studies. Lancet Planet Health. 2019;3:e469–e477.
20. Ibrahim MR, Haworth J, Cheng T. Understanding cities with machine eyes: a review of deep computer vision in urban analytics. Cities. 2020;96:102481.
21. Dubey A, Naik N, Parikh D, Raskar R, Hidalgo CA. Deep Learning the City: Quantifying Urban Perception At A Global Scale. *arXiv:160801769 [cs]* 2016. http://arxiv.org/abs/1608.01769 (accessed 3 Oct2017).
22. Seresinhe CI, Preis T, Moat HS. Quantifying the impact of scenic environments on health. Sci Rep. 2015;5:16899.
23. Larkin A, Gu X, Chen L, Hystad P. Predicting perceptions of the built environment using GIS, satellite and street view image approaches. Landsc Urban Plan. 2021; 216:104257.
24. Duncan GE, Avery AR, Strachan E, Turkheimer E, Tsang S. The Washington State Twin Registry: 2019 Update. Twin Res Hum Genet. 2019;22:788–93.
25. Markevych I, Schoierer J, Hartig T, Chudnovsky A, Hystad P, Dzhambov AM, et al. Exploring pathways linking greenspace to health: Theoretical and methodological guidance. Environ Res. 2017;158:301–17.
26. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid Scene Parsing Network. *arXiv:161201105 [cs]* 2016. http://arxiv.org/abs/1612.01105 (accessed 16 Oct2017).
27. Minka T, Cleven R, Zaykov Y. TrueSkill 2: An improved Bayesian skill rating system. 2018;: 24.
28. US EPA O. Smart Location Mapping. 2014.https://www.epa.gov/smartgrowth/smart-location-mapping (accessed 21 Mar2022).
29. Geographic Information Systems & Science - County Level Urban Sprawl Indices. https://gis.cancer.gov/tools/urban-sprawl/ (accessed 27 Apr2022).
30. Area Deprivation Index for the US. https://www.neighborhoodatlas.medicine.wisc.edu/ (accessed 15 Apr2022).
31. Seresinhe CI, Preis T, Moat HS. Using deep learning to quantify the beauty of outdoor places. R Soc Open Sci. 2017;4:170170.
32. Wang R, Liu Y, Lu Y, Zhang J, Liu P, Yao Y, et al. Perceptions of built environment and health outcomes for older Chinese in Beijing: A big data approach with street view images and deep learning technique. Computers, Environ Urban Syst. 2019;78:101386.
33. Liu M, Han L, Xiong S, Qing L, Ji H, Peng Y. Large-Scale Street Space Quality Evaluation Based on Deep Learning Over Street View Image. 2019, pp 690–701.
34. Chen C, Li H, Luo W, Xie J, Yao J, Wu L, et al. Predicting the effect of street environment on residents' mood states in large urban areas using machine learning and street view images. Sci Total Environ. 2022;816:151605.
35. Javanmardi M, Huang D, Dwivedi P, Khanna S, Brunisholz K, Whitaker R, et al. Analyzing associations between chronic disease prevalence and neighborhood quality through google street view images. IEEE Access. 2020;8:6407–16.
36. Robinson O, Tamayo I, de Castro M, Valentin A, Giorgis-Allemand L, Hjertager Krog N, et al. The urban exposome during pregnancy and its socioeconomic determinants. Environ Health Perspectives. 2018;126:077005.
37. Hystadl P. Residential greenness and birth outcomes: separating effects from spatially correlated built environment factors. Environ Health Perspectives. 2013; 122:1095–1102.
38. Vlaanderen J, de Hoogh K, Hoek G, Peters A, Probst-Hensch N, Scalbert A, et al. Developing the building blocks to elucidate the impact of the urban exposome on cardiometabolic-pulmonary disease. Environ Epidemiol. 2021;5:e162.
39. Guillien A, Cadiou S, Slama R, Siroux V. The exposome approach to decipher the role of multiple environmental and lifestyle determinants in Asthma. Int J Environ Res Public Health. 2021;18:1138.
40. Wild CP. Complementing the Genome with an "Exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev. 2005;14:1847–50.
41. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012;41:24–32.

ACKNOWLEDGEMENTS

The authors are grateful for the technological and academic contributions of Google Street View, Microsoft TrueSkill, MIT Place Pulse, and PSPNet researchers which made the described work possible.

AUTHOR CONTRIBUTIONS

PH conceptualized and led the overall stud. AL developed and led the collection of perceptions, deep learning models and analyses. GED leads the WSTR study and contributed to interpretation of results and manuscript preparation. OA contributed to GIS analysis and interpretation of results and manuscript preparation. ARA coordinated the study with the WSTR and contributed to interpretation of results and manuscript preparation.

FUNDING

Research reported in this publication was supported by the National Institute Of Environmental Health Sciences of the National Institutes of Health under Award Number R21ES029722. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COMPETING INTERESTS

The authors declare no competing interests.

ETHICAL APPROVAL

The study was approved by the Oregon State University Institutional Review Board (IRB-2019–0264) and the Washington State University Institutional Review Board (IRB-14514).

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41370-022-00489-8.

Correspondence and requests for materials should be addressed to Perry Hystad.