Health & Place 73 (2022) 102733

Contents lists available at ScienceDirect

Health and Place

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

Development and validation of automated microscale walkability audit method

Bon Woo Koo *, Subhrajit Guhathakurta, Nisha Botchwey

School of City and Regional Planning, College of Design, Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Reliability
Automated audit
Computer vision
Google Street View images

ABSTRACT

Measuring microscale factors of walkability has been labor-intensive and expensive. To reduce the cost, various efforts have been made including virtual audits (i.e., manual audits using street view images) and the introduction of computer vision techniques. Although studies have shown that virtual audits (i.e., manual audits using street view images) can reliably replicate in-person audits, they are still prohibitively expensive to be applied to a large geographic area. Past studies used computer vision techniques to help automate the audit process, but off-the-shelf models cannot detect some of the important microscale walkability characteristics, falling short of fully capturing the multi-facetted concept of walkability. This study is one of the earliest attempts to use the combination of custom-trained computer vision models, geographic information systems, and street view images to automatically audit a complete set of items of a validated microscale walkability audit tool. This study validates the reliability of the automated audit with virtual audit results. The automated audit results show high reliability, indicating automated audit can be a highly scalable and reliable replacement of virtual audit.

1. Introduction

Walkability indices, such as Walk Score (Walk Score, n.d.) and EPA's walkability index (U.S. Environmental Protection Agency, 2015), have been predominantly based on macroscale factors such as land use mix and residential density. While macroscale factors of walkability are important for walkability, they are only a part of the multi-dimensional concept of walkability (Alfonzo, 2005). Recently, microscale factors of walkability, such as fine-grained urban design details and their maintenance quality, have gained attention among urban planning and public health researchers as their importance on walking behavior has gained increasing theoretical as well as empirical support in the literature on the relationship between the built environment and walking (Alfonzo, 2005; Cain et al., 2014; Sallis et al., 2015). Many walkability audit tools that focus on microscale factors have been developed and validated (for a comprehensive review of walkability measures, see Brownson et al. (2009)). For example, the Microscale Audit of Pedestrian Streetscapes (MAPS) has included fine-grained design details and maintenance quality of streetscape components.[1] Studies using MAPS have reported that microscale characteristics of the built environment have significant associations with various types of walking and active transport (Cain et al., 2014; Sallis et al., 2015). Furthermore, microscale factors are relatively easy, quick, and inexpensive to modify, making timely interventions for promoting active transport and physical activity much more feasible than macroscale factors.

Despite these strengths, microscale factors have been rarely incorporated into widely used walkability indices such as Walk Score because its measurements have heavily relied on on-site, manual audits or surveys. With the introduction of street view image services such as Google Street View, many studies examined the possibility of replicating in-person audits with virtual audits (i.e., audits that are done by human auditors using the Google Street View (GSV) service to reduce resources required to travel to the target streets). These studies generally reported high agreement between in-person and virtual audits and suggested that virtual audit using GSV images is a reliable method for auditing the built and social environment (Clarke et al., 2010; Griew et al., 2013; Gullón et al., 2015; Rundle et al., 2011).

However, although virtual audit can eliminate the time and resources required for in-person auditors to travel to the target streets physically, they still require a similar amount of audit time per item (Rundle et al., 2011) or even more time than in-person audit (Gullón et al., 2015). Assuming 10–15 min of audit time for each street segment

* Corresponding author. School of City and Regional Planning, 245 Fourth St. NW, Suite 204, Atlanta, GA, 30332, USA.
E-mail addresses: bkoo34@gatech.edu (B.W. Koo), subhro.guha@design.gatech.edu (S. Guhathakurta), nisha.botchwey@design.gatech.edu (N. Botchwey).

[1] https://drjimsallis.org/measure_maps.html.

https://doi.org/10.1016/j.healthplace.2021.102733
Received 5 September 2021; Received in revised form 6 December 2021; Accepted 7 December 2021
Available online 16 December 2021

(Griew et al., 2013), virtually auditing a major U.S. city at street-level is still prohibitively expensive and labor-intensive. This limited scalability indicates that virtual audits may not save enough time and resources to make the inclusion of microscale factors into widely used walkability indices feasible. Even with virtual audits, the identification of streets that need planning and design interventions remains challenging, and the geographic and temporal scope of research is limited.

Recent studies have shown that computer vision can efficiently and reliably extract the physical characteristics of streetscapes from street view images (Ki and Lee, 2021; Li et al., 2018a; Li et al., 2018b; Tang and Long, 2018; Wang et al., 2019). Using computer vision techniques, these studies have extracted some key qualities of streetscapes, such as how much greenery, building, and sky can be seen from street view images. However, these attempts have been limited by the use of pre-trained computer vision models, which are usually confined to detecting pre-defined lists of general objects. These lists often do not include many objects relevant to pedestrian experience. For example, a pre-trained model of Pyramid Scene Parsing Network by Zhao et al. (2017) is trained using ADE20K dataset that has 150 object classes (Zhou et al., 2017), and many items in MAPS are not included in the 150 classes. As a result, the methods in previous attempts can be considered insufficient in capturing the multi-dimensional attributes that mediate walkability.

This study explores the feasibility of developing a method for automatically replicating the complete set of items of a validated walkability audit tool using computer vision and street view images. An important part of this study is the validation of the tool developed, which is undertaken by comparing the results from the automated audit with a virtual audit done by a trained human auditor.

2. Literature review

Around 2010, about three years after the launch of the Google Street View image service, researchers in urban planning and public health started to examine the potential of GSV images in replacing in-person audits. The majority of these studies focused on evaluating the agreement between the in-person audit and virtual audit. The findings from the studies showed that virtual audit can reliably replicate the in-person audit with some caveats. Clarke et al. (2010) compared in-person audit and virtual audit of 244 streets in Chicago, IL using a 29-item audit tool. They reported that indicators of "the built environment and neighborhood social and physical disorder" were reliably audited using the street view images. The observed agreement and kappa statistic (κ) were particularly high for objectively observable items such as "signs advertising alcohol (observed agreement = 0.92, κ = 0.34) or the presence of trees lining the street (observed agreement = 0.94, κ = 0.49)" (p.1227). However, the agreement was lower for items that need finer observations, such as the presence of garbage, litter, or broken glass (observed agreement = 0.35, κ = 0.04). They explained that there is a five-year time difference between the in-person audit and the virtual audit, and these items are "… likely to have changed substantially over the five years between the in-person and virtual audit" (p. 1227). They noted the agreement between the in-person and virtual audit was comparable to the inter-rater reliability between in-person audits. They conclude that "… some of the variability in characteristics observed across modes of observation may in fact be due to inter-rater reliability or test-retest reliability over the five years between observations" (p. 1228).

Other studies generally shared the finding that virtual audits can reliably replicate in-person audits except for ephemeral items. Badland et al. (2010) conducted a similar study in New Zealand, comparing in-person and virtual audits on 48 streets using a 21-item audit tool. They found that in-person and virtual audit showed acceptable levels of agreement on the majority of the items with intraclass correlation coefficient ≥0.70. A few items, including the number of fixed traffic controls, neighborhood permeability, and land use mix, showed agreements below the acceptable agreement level. Similar to Clarke et al. (2010), there were about two years of time lag between the in-person and virtual audit, which may have contributed to the low agreement for ephemeral items (e.g., litters on sidewalks). Rundle et al. (2011) conducted a similar comparison between in-person and virtual audit on 38 high-walkability block segments in New York City, NY. Among 103 categorical measures, 82.5% of the items show moderate or higher agreement (i.e., agreement ≥0.60). Among 37 count or proportion items, 62.1% of the items showed moderate or higher correlation (i.e., Spearman rank-order correlation ≥0.40). The agreement and correlations were higher for large or less temporally variable items. One important note in Rundle et al. (2011) is that because most street view images are taken by cameras attached on top of cars, there usually is some distance between the location of cameras and the sidewalk. Small items can be difficult to discern, particularly when they are located on the sidewalk surface, low to the ground, as they can easily be hidden behind other objects. Another study done by Griew et al. (2013) showed a similar result from 54 streets in the UK, finding that "percent agreement between in-person and desk-based audits … was high across all street characteristic categories with results ranging from 75 to 97% agreement (average 84%) and the kappa co-efficient ranging from k = 0.5 to 0.9 (moderate to almost perfect)" (p.5). They noted that the inter-rater reliability varied substantially between land uses, with the lowest agreement found in industrial areas and the highest found in residential areas. Similar to all other studies, they explained that items requiring "a judgment on quality or aesthetics" are commonly found to have low reliability.

The audit measurements using street view images are associated with various other outcome variables linked with the built environment condition. The outcome variables shown to have associations include observed physical activity (Kelly et al., 2014), children's antisocial behavior (Odgers et al., 2012), the severity of pedestrian crashes (Hanson et al., 2013), gentrification (Hwang and Sampson, 2014), neighborhood disorder (Bader et al., 2017; Marco et al., 2017), and violent crime (He et al., 2017). These studies further support the validity of using street view images for auditing the built as well as social environment.

The virtual audits used in most of these studies were, however, still manually done. For a fully automated extraction of built environmental information from street view images, recent studies started utilizing computer vision algorithms for detecting various streetscape objects (Koo et al., 2021). Although this approach is proven to be effective in capturing some of the pedestrian-related streetscape objects, it is often limited by what the pre-trained, off-the-shelf computer vision models offer. Many of these pre-trained models are trained using image datasets such as ImageNet (Deng et al., 2009), Microsoft COCO (Lin et al., 2014), and ADE20K (Zhou et al., 2017). The object classes included in these datasets do not contain many objects that are important to pedestrian experience (e.g., buffers and curb ramps) and their quality. While walkability is a composite concept that involves various factors from macroscale to microscale, many past studies are therefore limited to a few proxies of walkability (e.g., green view index or sky view factor) and rarely included the full suite of objects that collectively determine walkability. To the best of the author's knowledge, few studies, if any, have used fully automated methods (e.g., GIS or computer vision) to audit the full suite of items in validated walkability audit tools such as the Microscale Audit of Pedestrian Streetscapes (MAPS) or the Neighborhood Environment Walkability Survey (NEWS).

3. Data and analytical methods

Reliably replicating virtual audits with automated audits by computer vision and GIS is a multi-stage process. These stages include the selection of audit tools, the selection and training of computer vision models, the collection and processing of street view images, the training of a human auditor, and assessing the agreement of audits by the human auditor and computer vision models.

3.1. Audit tool

This study uses a shortened version of the Microscale Audit of Pedestrian Streetscapes (MAPS-mini) for the comparison between manual and automated audits (Sallis et al., 2015). The MAPS-mini is a 15-item version of the full, 120-item MAPS tool developed by Sallis et al. (2015). This audit tool is chosen for this study because (1) it has been validated to have statistically significant associations with active transport such as walking and biking (Sallis et al., 2015), (2) its short design makes the development of a training dataset for computer vision models more feasible, (3) the 15-item version maintains strong validity despite its short length, with a high correlation (r = 0.85) with the 120-item version of MAPS, (4) it contains relatively fewer items that require subjective judgment compared to, for example, NEWS, and (5) the items included in the audit are relatively stable over time and do not tend to change rapidly.

The MAPS-mini consists of two parts: The Crossing and the Segment. As shown in Table 1 below, the Crossing part has questions on three items, including walk signals, curb ramps, and marked crosswalks. The Segment part contains questions on street designs and qualities. The total point for each segment is calculated by summing the point for each item.

Some adjustments are made for more efficient use of computer vision models. First, some items in the Segment part are audited using geographic information systems (GIS) rather than computer vision models, as they can easily, and perhaps more effectively, be audited using conventional GIS data. Many municipalities and government entities maintain GIS databases on land use and transportation infrastructures (e.g., bike facilities and transit stops), and this study uses such databases to audit the following items in MAPS-mini: the primary type of land use (i.e., residential or commercial?) and presence of parks, transit stops, and bike lanes. The second adjustment is that, while the original instruction for the MAPS-mini direct users to audit *only one of two crossings* of any given street segment, this study audits crossings on both ends of a segment separately. The original MAPS-mini is designed to be applied to all street segments in a given area or all street segments in a route between origin and destination. In such cases, auditing only one of two crossings can be sufficient, as it is guaranteed that the other crossing will be considered when the next street segment is audited. However, this study randomly sampled streets for validation, and the sampled streets are not contiguous, which requires that both crossings be audited. A modification corollary to this adjustment is that this study simplifies the curb ramp question (i.e., the second item in the Crossing part) by deleting the third option, "Yes, at both pre-crossing and post-crossing curb(s)." Because this study audits two crossings at both ends of a segment separately, this option is redundant. Third, while the streetlights in the MAPS-mini is categorical, this study records the count of streetlights because categorizing the count of streetlights into None, Some, and Ample is ambiguous, and creating an objective criterion for the categorization is challenging. As the categorization of streetlights can be obtained post hoc from the count, this modification is not a loss of information. Finally, the question "What percentage of the length of the sidewalk/walkway is covered by trees, awnings or other overhead coverage?" is modified to (1) consider shade on both sidewalk and non-sidewalk areas altogether and (2) exclude awnings or other overhead coverage from consideration and only consider tree coverages.

Table 1
The original items on MAPS-mini and the method for automated audit. Total point is calculated by summing the scores in parenthesis across all items.

Part	Item	Audit Method
Crossing	Is a pedestrian walk signal present? *Possible answers: No(0)/Yes(1)* Is there a ramp at the curb(s)? Possible answers: No(0)/Yes (1) Is there a marked crosswalk? *Possible answers: No(0)/Yes(1)*	Computer vision
Segment	Type of land use? *Possible answers: Residential(0)/Commercial(1)* How many public parks are present? *Possible answers: 0(0)/1(1)/2 or more(2)* How many public transit stops are present? *Possible answers: 0(0)/1(1)/2 or more(2)* Is there a designated bike path? *Possible answers: No(0)/Painted line(1)/Physical Barrier (2)*	GIS
	Are there any benches or places to sit (include bus stop benches)? *Possible answers: No(0)/Yes(1)* Are streetlights installed? *Possible answers: None(0)/Some(1)/Ample(2)* Are the buildings well maintained? *Possible answers: 0–99%(0)/100%(1)* Is graffiti/tagging present (do not include murals)? *Possible answers: No(0)/Yes(1)* Is a sidewalk present? Possible *answers: No(0)/Yes(1)* Are there poorly maintained sections of the sidewalk that constitute major trip hazards? (e.g., heaves, misalignment, cracks, overgrowth, incomplete sidewalk) *Possible answers: None(0)/Any or No sidewalk present(1)* Is a buffer present? *Possible answers: No or No sidewalk present(0)/Yes(1)* What percentage of the length of the sidewalk/walkway is covered by trees, awnings, or other overhead coverage? *Possible answers: 0–25% or no sidewalk(0)/26–75%(1)/76–100%(2)*	Computer vision

3.2. Street view images

For the accurate representation of the streetscapes, the street view images need to be systematically collected to ensure that they cover all views of a given street segment that are relevant to the MAPS-mini. All street view images are downloaded through Google Street View API and are 640 by 640 pixels large. The field of view (FOV), a parameter that determines the horizontal field of view of the image, is set to 90.

This study collects images for the Crossing part and the Segment part of MAPS-mini separately. For each street segment, which is defined as a continuous stretch of a street defined by two intersection points at either end, there are two intersections, and two images are downloaded for each intersection, resulting in four intersection images for each street segment (heretofore, Crossing images). Two images per intersection are needed because there are many streets that are wider than what a street view image with 90° FOV can capture. The headings of the images are calculated using the sf package in R 4.0.2 in the following ways: First, using the road centerline shapefile from the Topologically Integrated Geographic Encoding and Referencing database (TIGER), the heading from the first (last) vertex of a street segment to the second (second to the last) vertex is calculated, namely the straight heading. Second, for each intersection, two images with headings equal to 'straight heading – 45' and 'straight heading +45', respectively, are downloaded.

For the Segment part, the exact locations of Google Street View images are unknown a priori. First, the metadata of street view images is downloaded at every 5 m to identify the exact locations of as many existing street view images as possible (see Fig. 1). Once the locations of images are identified, the headings for each of those image locations are calculated such that there are two images for each location that are perpendicular to the street segment and are looking back to back (i.e., looking at both sides of the segment). Finally, images looking up are downloaded for all of the image locations, creating a 'virtual tunnel' of images.

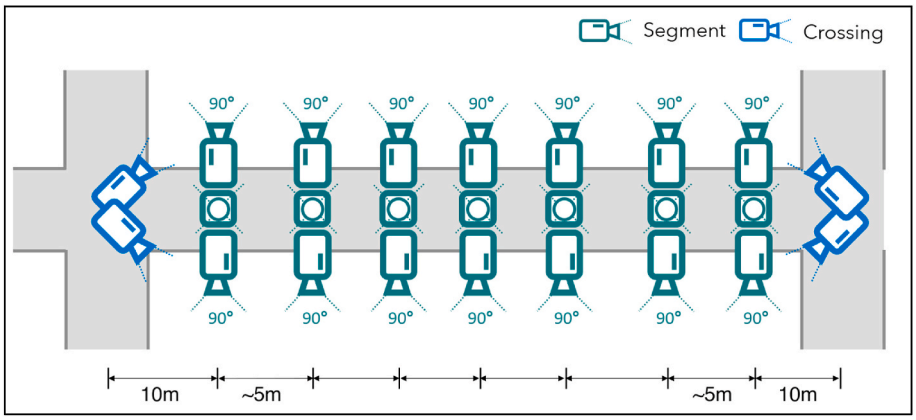

Fig. 1. Location and heading of street view images.

3.3. Computer vision

This study mainly uses Mask R–CNN,[2] an instance segmentation architecture by He et al. (2017), for its ability to (1) detect various objects within an image, (2) tell apart one object from others even among objects of the same kind (needed to count the occurrence of, e.g., streetlights), and (3) provide pixel-level masks for each object.

Although deep learning techniques applied to image data have shown remarkable success in extracting valuable information, their effectiveness is usually dependent on the size of the training dataset (Jean et al., 2016). For example, He et al. (2017) trained Mask R–CNN using the Microsoft COCO (Common Objects in Context), a large dataset for training computer vision models that contains 165,482 train, 81,208 validation, and 81,434 test images (Lin et al., 2014), to demonstrate the performance of their architecture and to distribute pre-trained weights. The challenge of this study was that there is no existing labeled dataset that contains all items on the MAPS-mini and that creating as large a dataset as the Microsoft COCO is infeasible. Although past implementations of computer vision architectures indicate that they may be trained to detect items on the MAPS-mini, the lack of a large training dataset makes the application of these techniques challenging.

To overcome this challenge, this study uses a technique called 'transfer learning.' Transfer learning loads existing weights that are trained on a large dataset, such as Microsoft COCO or ImageNet, on the model architecture. Next, the layers that are responsible for the prediction of, in the case of Mask-RCNN as an example, class categories, bounding boxes (i.e., rectangular markers that denote the location of objects in an image), and masks (i.e., pixel-wise overlays for the exact shape of the boundary of objects in an image) are replaced with new layers with untrained weights (see Fig. 2 for illustrations). In the training process through transfer learning, only these new layers are trained; all other layers which are loaded with weights trained on a large dataset are frozen in the training process to take advantage of their ability for image feature extraction. This training method significantly reduces the number of parameters that need to be trained and allows users to 'borrow' the performance of a model trained on a large dataset and repurpose the model for a different task with a much smaller training dataset.

This study trains three separate models for Segment part images, Crossing part images, and the vertical view of Segment part images for the ease of labeling the training dataset. Example images of the labeled training images are shown in Fig. 3. The number of images labeled for the training and validation of computer vision models is 2,000 and 500, respectively, with 20 classes for the Segment part of the MAPS-mini, about 420 and 110 images, respectively, with nine classes for the Crossing part, and 700 and 170 images, respectively, with three classes for the vertical view of Segment part. Street view images from random locations in Atlanta were sampled to create the training dataset. The random locations did not include street segments used for the validation. Because some items on the audit tools are rarely observed in the training dataset, a few non-street view, generic images on building maintenance conditions and trip hazards were acquired through the internet search into the training dataset. In the training process, most of the default hyperparameters provided in the public repository (https://github.com/matterport/Mask_RCNN) were used without modification except for a few minor changes to reflect the characteristics of street view images and the objects of interest, including number of classes, minimum image dimension, and anchor aspect ratios of the region proposal network. Image augmentation was heavily used to compensate for the small training dataset. The code needed to acquire street view images and apply Mask R–CNN with the trained weight files is available on a public repository.[3]

In addition to Mask R–CNN, this study also uses Pyramid Scene Parsing Network[4] (PSPNet), a semantic segmentation architecture, which generates predictions of what each pixel is likely to represent, to answer, "What percentage of the length of the sidewalk/walkway is covered by trees, awnings, or other overhead coverage?" PSPNet pre-trained on ADE20K can detect trees at the pixel level. PSPNet was applied to the images that are looking vertically up to estimate the proportion of the image covered by trees.

3.4. Geographic information systems

Four items in the MAPS-mini are audited using GIS and publicly available datasets. The GIS shapefiles of zoning designation, public parks, public transit stops, and bike paths were downloaded from Atlanta Regional Commission's data portal. The main land use is determined by intersecting 15-m buffer of street segments with the zoning shapefile. If more than 50% of the intersected area is commercial uses, the segment is classified as commercial. If otherwise, it is classified as residential. Public parks and transit stops are considered to be adjacent to a given segment if they fall into 15-m buffer of the segment. For bike paths, midpoints of segments are generated, and 5-m buffers of the midpoint are created. If there is a bike path intersecting with the buffer, the type of the intersecting bike path (e.g., painted line or physical barrier between bike land and road) is assigned to the segment.

[2] https://github.com/matterport/Mask_RCNN.

[3] https://github.com/BonwooKoo/Auto_MAPS.

[4] https://github.com/Vladkryvoruchko/PSPNet-Keras-tensorflow.

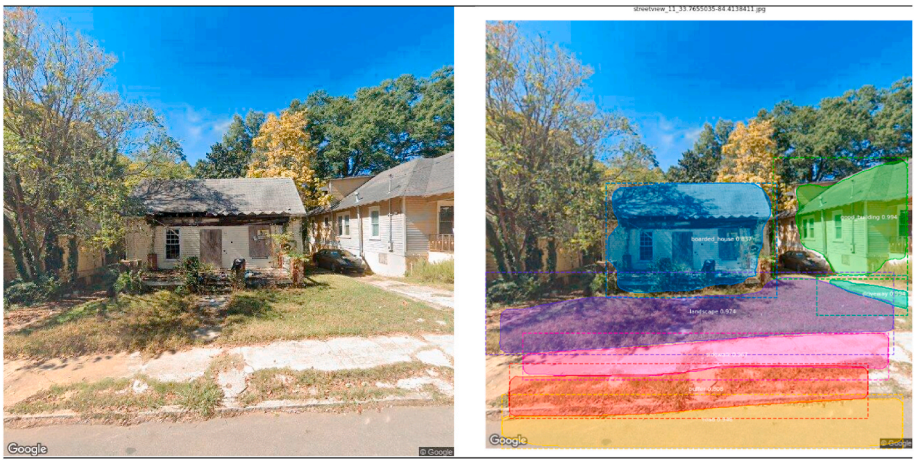

Fig. 2. Examples of class categories, bounding boxes, and masks generated from Mask R–CNN.

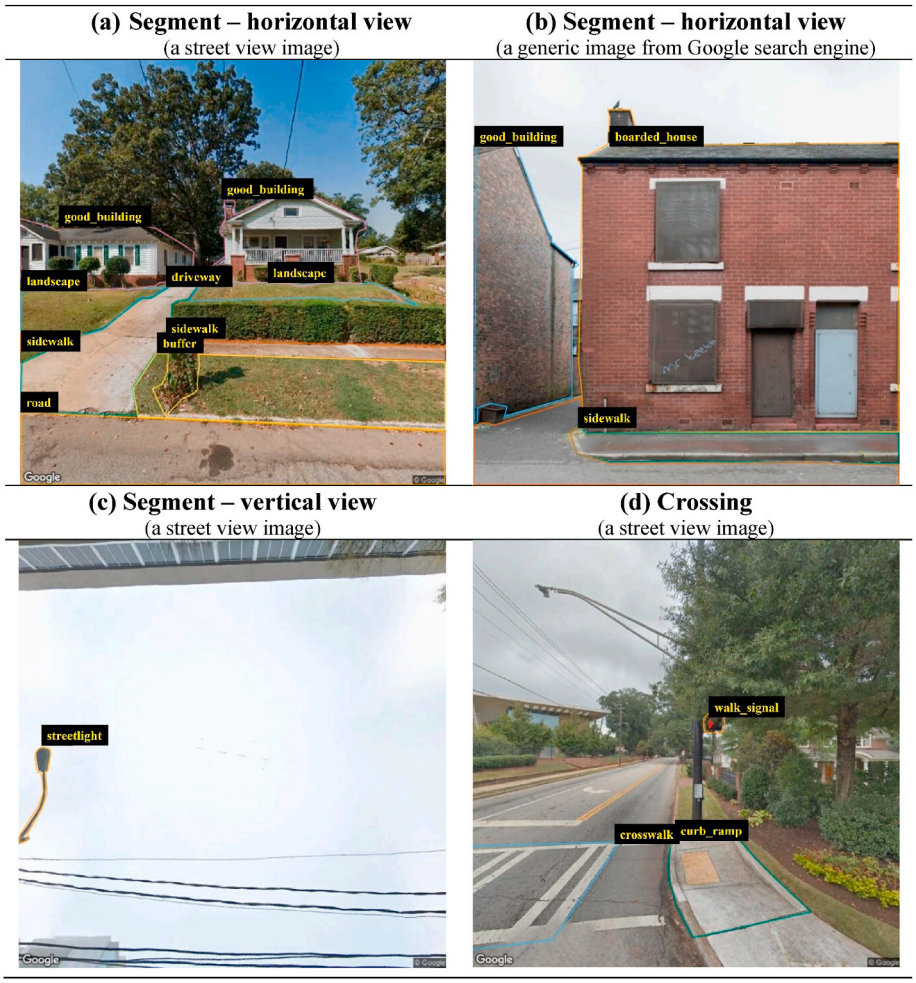

Fig. 3. Example images of the labeled training data.

3.5. Validation

This study uses stratified random sampling to select 100 street segments in Atlanta, GA, for validation. The size of this sample is determined based on past studies that assessed the reliability of audit tools on similar streetscape items (Badland et al., 2010; Clarke et al., 2010; Griew et al., 2013; Rundle et al., 2011). The four strata are defined by Walk Score and poverty rate at the Census Tract level using 2018 American Community Survey. The Walk Score is used to reflect the fact that in U.S. cities, typical streetscapes in urban and suburban settings (e.g., as measured by Walk Score) tend to be distinct. The quality of microscale streetscapes can also vary depending on poverty rate even among areas with similar macroscale walkability (Bereitschaft, 2017; Neckerman et al., 2009). Using the median values of Walk Score and poverty rate, the strata are defined as 'High Walk Score – High poverty rate,' 'High Walk Score – Low poverty rate,' 'Low Walk Score – High poverty rate,'

and 'Low Walk Score – Low poverty rate.' Proportional to the relative counts of Census Tracts that fall into each stratum, the total of 100 street segments are randomly selected. Fig. 4 shows the location of four strata used for sampling streets as well as the location of the selected streets. Before the actual audit, two test auditors were recruited to conduct a pilot test of MAPS-mini audit on 15 street segments. Feedback from the pilot test auditors contributed to the development of an audit guideline, which was used to train the primary auditor. This auditor used Google Maps to audit 100 selected street segments using all functionalities available, including zooming and panning.

The images for the same streets were processed using the computer vision models and GIS. Before calculating how well the results from virtual audit and automated audit agree with each other, the information from computer vision models and GIS needed to be aggregated to street-level because, as illustrated in Fig. 1, there are multiple images for each street. To do so, the count of the item of interest is summed up for each street segment and converted into respective categorical answers. For example, "How many public transit stops are present?" is answered by summing up the number of transit stops for a given street and converting the number into "0", "1", "2 or more" categories. For sidewalks and buffers, a given street segment is considered to have a sidewalk when sidewalks are detected in at least two consecutive images on the segment. Buffers detected in images without sidewalks are not counted. Because the MAPS-mini is about the pedestrian experience, only trip hazards that are detected with sidewalks are counted. The percentage of the sidewalk covered by overhead trees is calculated by converting the numeric output from PSPNet into discrete categorical answers. First, the proportion of pixels representing trees is calculated from PSPNet output. Next, if the proportion of trees is less than 10% or if there is no sidewalk at all, the lowest tree coverage category is assigned. If the proportion of pixels representing trees is between 10% and 20%, the medium tree coverage category is assigned. If else, the high tree coverage category is assigned. These values of 10% and 20% are different from the original MAPS-mini answer options (i.e., 25% and 75%) because the original answer options are the percentage of tree coverage *above the sidewalks* while the proportion of trees acquired from PSPNet represents tree coverage of the entire vertical view.

The reliability of automated audit is assessed using percent agreement and Cohen's Kappa (McHugh, 2012). Streetlight is an exception, as this study modified the question of "Are streetlights installed?" to have a numeric answer rather than a categorical one. This study calculated intraclass correlation coefficient (ICC) to assess the agreement for this item using two-way mixed effects, absolute agreement, single rater model (Koo and Li, 2016). Statistical analysis is conducted in R 4.0.2. Cohen's Kappa and ICC are calculated using psych 2.0.8 package. The total point has a numeric value and is assessed using Pearson's correlation.

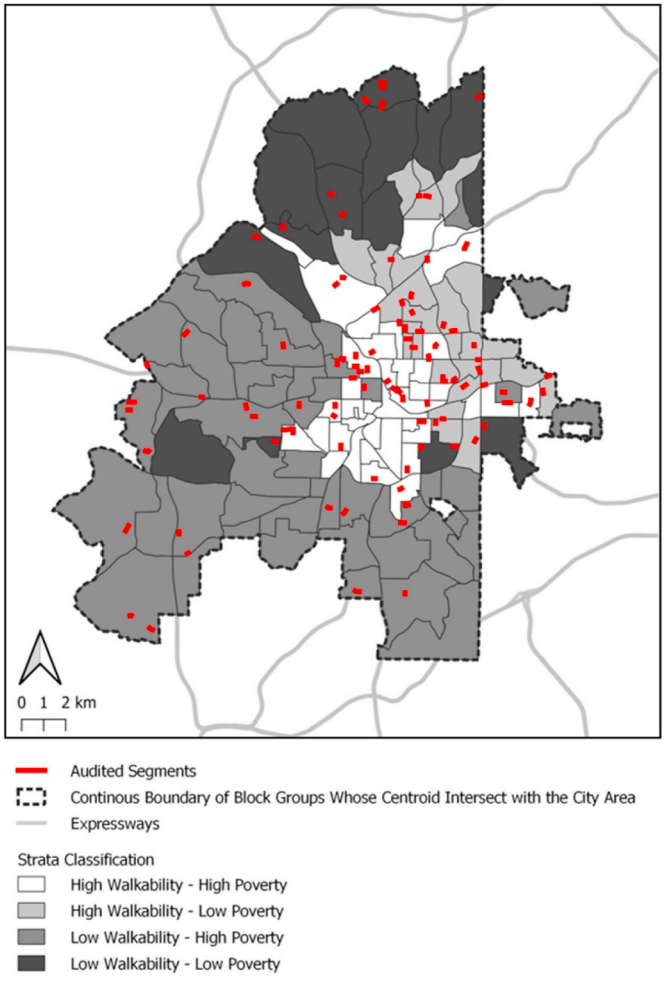

Fig. 4. Four strata by Walk Score and poverty rate overlayed with the selected street segments.

4. Results

Of the 100 street segments, four street segments had considerable mismatches in road configuration between TIGER shapefile and Google or were missing GSV images. These segments were excluded. The average length of the 96 streets is 125.1 m with standard deviation of 54.2 m. The street length ranged between 51.3 and 299.4 m. The street view images used in virtual and automated audits are temporally well-aligned: The virtual audit used images from 2016 to 2021 with the mean of 2019.28 and standard deviation of 0.790. The images used in automated audit are taken between 2015 and 2020 with a mean of 2019.12 and standard deviation of 0.847.

The total number of images used for the validation is 3,386, with about 35 images per street segment. Among the 3,386 images, 2,998 were for the Segment part, of which 996 are images looking upward, and 388 images for the Crossing part. Fig. 5 shows example images of the prediction results that the computer vision model generated, where the masked images in the first row show images for the Segment part of MAPS-mini in which a few items such as road, sidewalk, buffer, and well-maintained buildings are detected. The second row of the figure illustrates two images for the vertical view of Segment part, one for the detection of a streetlight from Mask R–CNN and the other for the pixel-wise segmentation from PSPNet for tree canopy detection. The third row of the figure shows images for the Crossing part with crossing, curb ramp, and walk signals detected. The mean average precision (mAP), a commonly used metric for the performance of computer vision models, with the intersection over the union (IoU) threshold of 0.5 is 0.582 for the Segment part, 0.668 for the Crossing part, and 0.815 for the vertical view of the Segment part (for more on mAP of the original Mask R–CNN article, see He et al. (2017); for more on the definition of mAP, see Section 4.2. of Everingham et al. (2010)).

Table 2 shows the observed agreement and Kappa statistics with 95% confidence interval between virtual and automated audits. Among the categorical items, observed agreements of the Segment items are high (>0.80) except for the presence of ill-maintained buildings and shade from overhead trees, indicating high reliability between most of the Segment items between virtual and automated audits. Kappa coefficients of the corresponding items are generally lower than observed agreement, especially for items that are less commonly observed (i.e., the presence of graffiti, seating, and trip hazards). The item with the lowest level of agreement is the presence of ill-maintained buildings (observed agreement = 0.771, $\kappa = 0.108$). The items audited using GIS consistently show high levels of agreement (observed agreement = 0.865–0.990, $\kappa = 0.620$–0.852). The ICC for the number of streetlights between the virtual and the automated audit is 0.564, indicating a

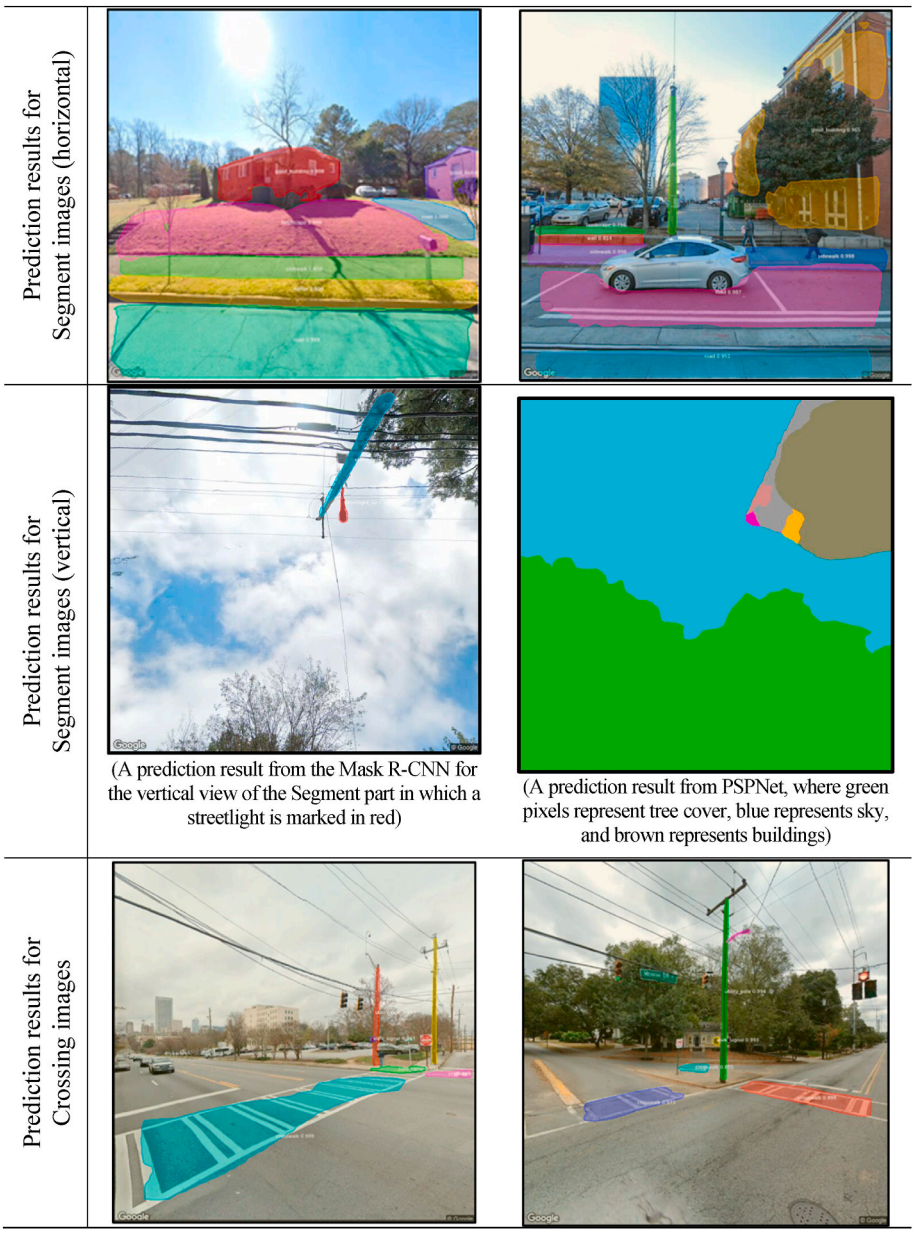

Fig. 5. Examples of prediction results from the computer vision models (bounding boxes are omitted from the display).

moderate correlation.

Similarly, the three Crossing part items show moderate to high observed agreements. The presence of walk signal for two intersections shows a high observed agreement with Kappa coefficient 0.777 and 0.897, respectively. The level of agreement for the presence of crosswalks and curb ramps indicates that these items are less reliably audited (observed agreement = 0.719–0.760, κ = 0.401–0.484). Finally, the correlation coefficient for the total points between virtual audit and automated audit is 0.739, indicating a high correlation.

Discussion and conclusion

Compared to virtual audit, this study shows that many audit items in the MAPS-mini can be reliably audited using the combination of computer vision, street view images, and GIS. Of 17 categorical items in Table 2, 11 items show high observed agreement (observed agreement >0.80), and most of the rest are of moderate agreement (observed agreement >0.60). Kappa coefficients tend to be lower than observed agreement, particularly for rare and/or qualitative items, such as the presence of ill-maintained buildings. The levels of agreement between virtual and automated audits are generally on par with the results of similar items from the past studies (e.g., Clarke et al., 2010; Griew et al., 2013). For some items, such as the presence of trip hazards, the levels of agreement are higher than similar items reported in the literature (e.g., 'street condition' with κ = 0.032 in Clarke et al. (2010)). It is demonstrated that computer vision combined with street view images has the potential to offer a reasonably reliable and highly scalable alternative to virtual audit at a significantly lower cost. The fact that this study used small training datasets for the computer vision algorithms suggests that the performance of automated audit has the potential to be significantly improved with bigger training datasets.

While the overall level of agreement is acceptable, some items have relatively low levels of Kappa coefficients. One potential reason for the low Kappa coefficients is the rare occurrences of those items in both the training dataset for the computer vision algorithms and the audited segments. There are only four seating and six graffiti found in the audited segments. This rarity can result in a high chance agreement, which can decrease the magnitude of Kappa coefficient (Sim and Wright,

Table 2
Levels of agreement between virtual and automated audit.

Class	Item (Presence of..)	Observed agreement or correlation	Kappa	CI
Segment	Buffer	0.833	0.658	0.509–0.807
	No graffiti	0.927	0.423	0.071–0.775
	Seating	0.958	0.314	−0.178–0.806
	Sidewalk	0.896	0.717	0.556–0.878
	No trip Hazard	0.823	0.379	0.141–0.617
	No ill-maintained Building	0.771	0.108	−0.102–0.318
	Shade from overhead tree	0.583	0.357	0.196–0.518
	Streetlight[a]	0.564	–	0.438–0.669
	Bike Path[b]	0.990	0.852	0.566–1.000
	Public Park[b]	0.948	0.764	0.574–0.953
	Contains Commercial Uses[b]	0.875	0.728	0.586–0.87
	Transit Stop[b]	0.865	0.620	0.489–0.752
Crossing 1	Walk Signal	0.979	0.897	0.758–1.000
	Crosswalk	0.760	0.481	0.300–0.662
	Curb Ramp	0.740	0.484	0.312–0.655
Crossing 2	Walk Signal	0.958	0.777	0.567–0.987
	Crosswalk	0.719	0.435	0.255–0.615
	Curb Ramp	0.719	0.401	0.220–0.582
	Total Point[c]	0.739	–	0.627–0.821

[a] Agreement for streetlight is measured using intraclass correlation based on single rater, absolute agreement.
[b] -way mixed-effects model.[b]These items are audited using GIS as there commonly exists public GIS dataset for these items.
[c] Agreement for total point is measured using Peason's correlation.

2005). For example, the chance agreement of seating is very high at 0.939. Plugging this value into the equation for the Kappa coefficient, which is (observed agreement – chance agreement)/(1 – chance agreement), it is apparent that this item requires an exceptionally high observed agreement to increase the Kappa coefficient. The rarity also means that these items appeared in training data less frequently, and the computer vision algorithm did not have enough to learn from.

Another reason for some of the low Kappa coefficient is rooted in the subjective nature of some item definitions. For example, the maintenance quality of buildings is intrinsically a continuous range with no objective cutoff line that separates good and bad maintenance quality. Labeling houses into either ill-maintained or normal buildings requires drawing an arbitrary cutoff line on the continuous scale. Despite the efforts to make the cutoff line as objective as possible by providing detailed examples in the audit guideline, this subjectivity is difficult to eliminate. This issue of subjectivity applies to graffiti (i.e., how much pleasing should it be aesthetically in order to qualify as a mural?) and trip hazard on sidewalks (i.e., how serious a damage should it be in order to be considered as a trip hazard?). Furthermore, trip hazards are not a single object but rather ill-maintained parts of sidewalks that can take a variety of appearances ranging from severely cracked sidewalks to overgrown grass that is tall enough to impede walking. This inconsistency in appearance can make detection more challenging. Walk signal is an example that illustrates the challenge posed by subjectivity and inconsistency in appearance. Because walk signals are objectively identifiable with virtually identical appearances in most cases and are clearly distinguishable from other objects that are not walk signals, distinguishing them is less ambiguous both for the human auditor and the computer vision algorithms, resulting in the high reliability score in Table 2.

Note that these issues of rarity and subjectivity are not intrinsic to computer vision algorithms and/or GSV images. The rarity of some items, such as graffiti or seating, is likely to be characteristics of the streetscapes specific to Atlanta and may not be attributable to GSV images. Also, past studies showed that items requiring subjective assessment tend to result in low levels of reliability even among human raters (Clarke et al., 2010; Clifton et al., 2007).

This study identified a few technical challenges associated with using computer vision algorithms and GSV images for streetscape audits. First, the maximum resolution of images that Google Street View API provides is 640 by 640 pixels. While this is a sufficient resolution when objects of interest are located close to the road from which photos were taken, objects that are located far from the road or small in size can appear blurry in GSV images. The issue of blurriness can be exacerbated by wide width of the roads. This blur can reduce the visual information of objects and potentially have negative impacts on the prediction accuracy of computer vision models. Second, the street view images are almost always taken with a camera attached on the top of cars. Because the vantage point of the camera is fixed to the roads, items on or closer to the ground (e.g., trip hazards, seating) can be challenging to audit using street view images because they can be hidden behind other objects such as parked cars and tree trunks (Rundle et al., 2011). Furthermore, the street view images are limited to public roads that Google can access.

The limitations of the study can provide insights on how future studies can improve upon this study. Due to the limitations in the available resources, increasing the size of the training data was infeasible. As one of the most important factors that gave a boost to deep learning-based computer vision algorithms is the appearance of large labeled datasets (Voulodimos et al., 2018), increasing training data to a sufficiently large size may provide substantial improvements to the effectiveness of automated audits, particularly for items that were rarely found in our data (i.e., graffiti and seating). Second, this study did not assess the reliability between in-person and automated audits. We chose to do so because past studies have demonstrated that virtual audit can reliably assess the streetscapes except for items that are temporally variable or small in size (Clarke et al., 2010; Rundle et al., 2011). Importantly, several studies have already used virtual audits as a sole source of information on the streetscapes without conducting in-person audits and linked the virtual audit result to outcome variable of their interest (e.g., Hanson et al., 2013; He et al., 2017; Mooney et al., 2016). The goal of this study is to develop an automated method that can replicate these virtual audits. Third, the labeling of the training dataset for the computer vision algorithm can be further refined. The maintenance quality of buildings, trip hazards, seating, and graffiti categories may benefit from a more detailed labeling strategy. For example, in addition to labeling an entire house as an 'ill-maintained building' or 'well-maintained building,' future studies may benefit from adding labels of individual components that can be collectively used to determine the building maintenance quality, such as boarded windows, cracks on outer structures, significantly overgrown landscapes, and substantial discoloration. Fourth, the selection of 100 street segments for the assessment of reliability may not be sufficient, particularly considering the rarity of some audited items (i.e., only four seating and six graffiti were found in the 100 streets). Future studies can benefit from carefully selecting the sample size with a consideration of this rarity. Fifth, it was found that there are many locations where the distance between two consecutive images is long (or short) enough to create a significant gap (or an overlap) between the two images, assuming that the FOV is set to 90°. The gaps and overlaps are frequently observed not only on curvy segments but also on completely straight segments. Future studies can benefit from careful processing that eliminates the gaps and overlaps before performing automated environmental audits. Finally, a few improvements are needed to make the automated audit method generalizable to other U.S. cities. Because Mask R–CNN used in this study was trained using images from Atlanta and is tested for Atlanta streets, the degree to which it can be applied to other cities is unknown. Furthermore, this study used GIS data to audit some items in the MAPS-mini. In many cities, GIS data on bike lanes and parks are either missing or unreliable. There is no standardized land use or zoning codes across cities, and auditing the primary type of land use through GIS may require additional pre-processing for each city. Important next steps for generalizability include expanding the training dataset for the computer

vision model to include a representative sample of U.S. cities and to replace GIS data with more generalizable data sources.

This study demonstrated the effectiveness of using GSV images and computer vision algorithms for automatically auditing the streetscapes. This study identified several challenges specific to Atlanta streetscapes and to GSV images, respectively, which could have been contributing factors to the lower levels of agreement observed for some items. However, despite these challenges, the results of this study demonstrate that computer vision algorithms and GSV images can provide a reliable method for automatically auditing streetscapes. If the challenges identified in this study are mitigated and overcome in future studies, the method is expected to offer a cost-effective method for conducting a streetscape audit using a scientifically validated audit tool (i.e., the MAPS-mini) at regional or even at national scale, a task that has been prohibitively expensive when using conventional methods.

Author note

We have no conflicts of interests to disclose.

References

Alfonzo, M.A., 2005. To walk or not to walk? The hierarchy of walking needs. Environ. Behav. 37 (6), 808–836. https://doi.org/10.1177/0013916504274016.

Bader, M.D.M., Mooney, S.J., Bennett, B., Rundle, A.G., 2017. The promise, practicalities, and perils of virtually auditing neighborhoods using Google street view. Ann. Am. Acad. Polit. Soc. Sci. 669 (1), 18–40. https://doi.org/10.1177/0002716216681488.

Badland, H.M., Opit, S., Witten, K., Kearns, R.A., Mavoa, S., 2010. Can virtual streetscape audits reliably replace physical streetscape audits? J. Urban Health 87 (6), 1007–1016. https://doi.org/10.1007/s11524-010-9505-x.

Bereitschaft, B., 2017. Equity in microscale urban design and walkability: a photographic survey of six Pittsburgh streetscapes. Sustainability 9 (7), 1233. https://doi.org/10.3390/su9071233.

Brownson, R.C., Hoehner, C.M., Day, K., Forsyth, A., Sallis, J.F., 2009. Measuring the built environment for physical activity: state of the science. Am. J. Prev. Med. 36 (4, Suppl. ment), S99–S123. https://doi.org/10.1016/j.amepre.2009.01.005 e12.

Cain, K.L., Millstein, R.A., Sallis, J.F., Conway, T.L., Gavand, K.A., Frank, L.D., Saelens, B. E., Geremia, C.M., Chapman, J., Adams, M.A., Glanz, K., King, A.C., 2014. Contribution of streetscape audits to explanation of physical activity in four age groups based on the Microscale Audit of Pedestrian Streetscapes (MAPS). Soc. Sci. Med. 116, 82–92. https://doi.org/10.1016/j.socscimed.2014.06.042.

Clarke, P., Ailshire, J., Melendez, R., Bader, M., Morenoff, J., 2010. Using Google Earth to conduct a neighborhood audit: reliability of a virtual audit instrument. Health Place 16 (6), 1224–1229. https://doi.org/10.1016/J.HEALTHPLACE.2010.08.007.

Clifton, K.J., Livi Smith, A.D., Rodriguez, D., 2007. The development and testing of an audit for the pedestrian environment. Landsc. Urban Plann. 80 (1), 95–110. https://doi.org/10.1016/j.landurbplan.2006.06.008.

Deng, J., Dong, W., Socher, R., Li, L., Li, Kai, Fei-Fei, Li, 2009. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88 (2), 303–338. https://doi.org/10.1007/s11263-009-0275-4.

Griew, P., Hillsdon, M., Foster, C., Coombes, E., Jones, A., Wilkinson, P., 2013. Developing and testing a street audit tool using Google Street View to measure environmental supportiveness for physical activity. Int. J. Behav. Nutr. Phys. Activ. 10 (1), 103. https://doi.org/10.1186/1479-5868-10-103.

Gullón, P., Badland, H.M., Alfayate, S., Bilal, U., Escobar, F., Cebrecos, A., Diez, J., Franco, M., 2015. Assessing walking and cycling environments in the streets of Madrid: comparing on-field and virtual audits. J. Urban Health 92 (5), 923–939. https://doi.org/10.1007/s11524-015-9982-z.

Hanson, C.S., Noland, R.B., Brown, C., 2013. The severity of pedestrian crashes: an analysis using Google Street View imagery. J. Transport Geogr. 33, 42–53. https://doi.org/10.1016/j.jtrangeo.2013.09.002.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: IEEE International Conference on Computer Vision, pp. 2961–2969. http://openaccess.thecvf.com/content_ICCV_2017/papers/He_Mask_R-CNN_ICCV_2017_paper.pdf.

He, L., Páez, A., Liu, D., 2017. Built environment and violent crime: an environmental audit approach using Google Street View. Comput. Environ. Urban Syst. 66, 83–95. https://doi.org/10.1016/J.COMPENVURBSYS.2017.08.001.

Hwang, J., Sampson, R.J., 2014. Divergent pathways of gentrification: racial inequality and the social order of renewal in Chicago neighborhoods. Am. Socio. Rev. 79 (4), 726–751. https://doi.org/10.1177/0003122414535774.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353 (6301), 790. https://doi.org/10.1126/science.aaf7894.

Kelly, C., Wilson, J.S., Schootman, M., Clennin, M., Baker, E.A., Miller, D.K., 2014. The built environment predicts observed physical activity. Front. Public Health 2, 52. https://doi.org/10.3389/fpubh.2014.00052.

Ki, D., Lee, S., 2021. Analyzing the effects of Green View Index of neighborhood streets on walking time using Google Street View and deep learning. Landsc. Urban Plann. 205, 103920. https://doi.org/10.1016/j.landurbplan.2020.103920.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropractic Med. 15 (2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012. PubMed.

Koo, B.W., Guhathakurta, S., Botchwey, N., 2021. How are neighborhood and street-level walkability factors associated with walking behaviors? A big data approach using street view images. Environ. Behav. https://doi.org/10.1177/00139165211014609, 00139165211014609.

Li, X., Ratti, C., Seiferling, I., 2018a. Quantifying the shade provision of street trees in urban landscape: a case study in Boston, USA, using Google Street View. Landsc. Urban Plann. 169, 81–91. https://doi.org/10.1016/j.landurbplan.2017.08.011.

Li, X., Santi, P., Courtney, T.K., Verma, S.K., Ratti, C., 2018b. Investigating the association between streetscapes and human walking activities using Google Street View and human trajectory data. Trans. GIS 22 (4), 1029–1044. https://doi.org/10.1111/tgis.12472.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, pp. 740–755.

Marco, M., Gracia, E., Martín-Fernández, M., López-Quílez, A., 2017. Validation of a Google street view-based neighborhood disorder observational scale. J. Urban Health : Bull. N. Y. Acad. Med. 94 (2), 190–198. https://doi.org/10.1007/s11524-017-0134-5.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Med. 22 (3), 276–282. PubMed.

Mooney, S.J., DiMaggio, C.J., Lovasi, G.S., Neckerman, K.M., Bader, M.D.M., Teitler, J. O., Sheehan, D.M., Jack, D.W., Rundle, A.G., 2016. Use of Google street view to assess environmental contributions to pedestrian injury. Am. J. Publ. Health 106 (3), 462–469. https://doi.org/10.2105/AJPH.2015.302978.

Neckerman, K.M., Lovasi, G.S., Davies, S., Purciel, M., Quinn, J., Feder, E., Raghunath, N., Wasserman, B., Rundle, A., 2009. Disparities in Urban neighborhood conditions: evidence from GIS measures and field observation in New York City. J. Publ. Health Pol. 30 (S1), S264–S285. https://doi.org/10.1057/jphp.2008.47.

Odgers, C.L., Caspi, A., Bates, C.J., Sampson, R.J., Moffitt, T.E., 2012. Systematic social observation of children's neighborhoods using Google Street View: a reliable and cost-effective method. JCPP (J. Child Psychol. Psychiatry) 53 (10), 1009–1017. https://doi.org/10.1111/j.1469-7610.2012.02565.x.

Rundle, A.G., Bader, M.D.M., Richards, C.A., Neckerman, K.M., Teitler, J.O., 2011. Using Google street view to audit neighborhood environments. Am. J. Prev. Med. 40 (1), 94–100. https://doi.org/10.1016/j.amepre.2010.09.034.

Sallis, J.F., Cain, K.L., Conway, T.L., Gavand, K.A., Millstein, R.A., Geremia, C.M., Frank, L.D., Saelens, B.E., Glanz, K., King, A.C., 2015. Is your neighborhood designed to support physical activity? A brief streetscape audit tool. Prev. Chronic Dis. 12 https://doi.org/10.5888/pcd12.150098.

Sim, J., Wright, C.C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. 85 (3), 257–268. https://doi.org/10.1093/ptj/85.3.257.

Tang, J., Long, Y., 2018. Measuring visual quality of street space and its temporal variation: methodology and its application in the Hutong area in Beijing. Landscape and Urban Planning. https://doi.org/10.1016/J.LANDURBPLAN.2018.09.015.

U.S. Environmental Protection Agency, 2015. National walkability index. https://www.epa.gov/smartgrowth/smart-location-mapping.

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep Learning for Computer Vision: A Brief Review. Computational Intelligence and Neuroscience, p. 7068349. https://doi.org/10.1155/2018/7068349, 2018.

Walk Score. Walk score methodology. https://www.walkscore.com/methodologyhtml.

Wang, R., Lu, Y., Zhang, J., Liu, P., Yao, Y., Liu, Y., 2019. The relationship between visual enclosure for neighbourhood street walkability and elders' mental health in China: using street view images. J. Trans. Health 13, 90–102. https://doi.org/10.1016/J.JTH.2019.02.009.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network, pp. 2881–2890.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ADE20K dataset. In: Proceedings Of the IEEE Conference On Computer Vision And Pattern Recognition (CVPR), pp. 633–641. http://groups.csail.mit.edu/vision/datasets/ADE20K/.