Highlights

A citizen science toolkit to collect human perceptions of urban environments using open street view images

Matthew Danish,S.M. Labib,Britta Ricker,Marco Helbich

- Studies using commercial street view imagery have proliferated despite licensing terms.
- We built a workflow and webapp to collect perceptions of open street view imagery.
- The webapp presents a simple and consistent interface with a swipe-to-rate UI.
- Our data preparation methods and mobile-friendly survey are open, FAIR and reusable.
- Anyone may easily clone, modify & deploy this perception survey in any desired place.

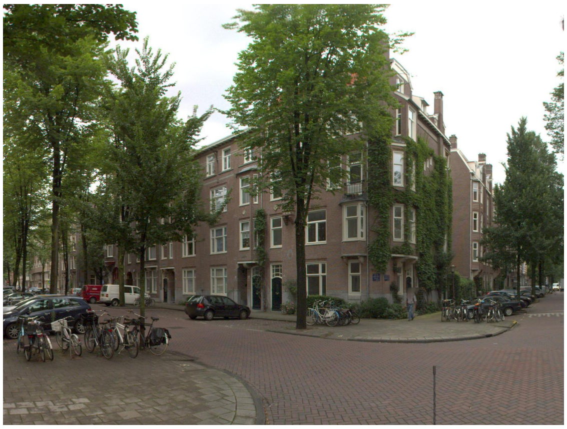


Figure 1: Sample image from Amsterdam (source: Mapillary)

A citizen science toolkit to collect human perceptions of urban environments using open street view images

Matthew Danish[*], S.M. Labib, Britta Ricker and Marco Helbich

Utrecht University, Princetonlaan 8a, Utrecht 3584 CB, The Netherlands

ARTICLE INFO

Keywords:
street view
open source
human perception
environment
toolbox
citizen science

ABSTRACT

Street View-level Imagery (SVI) is a valuable data source for studies (e.g., environmental assessments, green space identification or land cover classification). While commercial SVI is available, such providers commonly restrict copying or reuse in ways necessary for research. Open SVI datasets are readily available from less restrictive sources, such as Mapillary, but due to the heterogeneity of the images, these require substantial preprocessing, filtering, and careful quality checks. We present an efficient method for automated downloading, processing, cropping, and filtering open SVI, to be used in a survey of human perceptions of the streets portrayed in these images. We demonstrate our open-source reusable SVI preparation and smartphone-friendly perception-survey software with Amsterdam (Netherlands) as the case study. Using a citizen science approach, we collected from 331 people 22,637 ratings about their perceptions for various criteria. We have published our software in a public repository for future re-use and reproducibility.

1. Introduction

'Would you feel safe in this place? Does it look pleasant? Does it feel walkable?' These are some examples of questions that we can ask people about their view of a city from the street level. The answers will vary from person to person, and there is no single 'right answer'. Some people may have longstanding associations with a given place 'soaked in memories and meanings' (Lynch, 1960, p. 1), others may have never visited. The subjective answers that people give for certain built and natural environmental characteristics (e.g., walkability or greenness) might differ considerably from objectively measured indices (Kothencz and Blaschke, 2017; Lotfi and Koohsari, 2009). Although subjective perception varies between people, that does not make perception less important; on the contrary, 'a person's quality of life is dependent on the exogenous (objective) facts of his or her life and the endogenous (subjective) perception he or she has of these factors and of himself or herself' (Dissart and Deller, 2000, p. 136). Capturing data on

[*] Corresponding author
m.r.danish@uu.nl (M. Danish)
ORCID(s): 0000-0002-7186-387X (M. Danish)

subjective perception gives researchers the opportunity to study the correlation or differences between such exogenous facts and endogenous perceptions; for example, urban development researchers could investigate questions like: do crime statistics in urban areas correspond with perceived safety in those same places? Or, for sustainable transport researchers: is there actually more walking and/or cycling activity measured on streets that people perceive to be more walkable and/or bikeable? The answers to these questions could be used to guide future planning and development policy, leading to substantial improvement in people's perception of their built environment and subsequently their quality of life.

Nonetheless, most urban design, transportation planning and environmental epidemiology researchers rely on objectively measured streetscape indicators because it is easier to collect measurable and quantifiable physical attributes, such as for indicators of green spaces (Kothencz and Blaschke, 2017; Labib et al., 2020). However, people's subjective feelings about places do not necessarily correlate with such objectively measured spatial indicators (McCrea et al., 2006). To capture those sentiments as usable data, researchers typically use resource-intensive methods such as field interviews or questionnaires (Lynch, 1960; Lotfi and Koohsari, 2009). Therefore, we argue that developing an efficient yet easy-to-implement method to capture human perceptions of urban environments would provide a much needed tool for answering questions like those posed above. Taken further, these answers can help with efforts to achieve the United Nations Sustainable Development Goals[1] such as Goal 11, 'Sustainable Cities and Communities'.

Over the past decade, various companies, organisations and people have been collecting Street View-level Imagery (SVI). The resulting datasets of SVI have been successfully used in numerous studies over a wide variety of domains (Biljecki and Ito, 2021). Although SVI has gained momentum as a data source across disciplines, including perception collection, considerable challenges remain due to specific limitations of various SVI sources. In particular, we focus on two major classes of SVI currently accessible to researchers: (i) commercial SVI such as offered by Google, Baidu or Tencent and (ii) Volunteered Street View-level Imagery (VSVI) from open platforms such as Mapillary and KartaView (formerly OpenStreetCam), which collect geo-tagged volunteer-submitted photographs and make them available from a central repository.

Unlike commercial SVI platforms like Google, VSVI platforms such as Mapillary permit usage of their imagery under the Creative Commons Attribution-ShareAlike 4.0 International[2] license, making it easy to adapt into an open-source software system for collecting human perception responses. However, the quality of VSVI can be vary considerably from very poor to excellent (Ma et al., 2019). Recent work by Zheng and Amemiya (2023) has highlighted ways to improve the usability of VSVI by undergoing a process of filtering, however, no studies have yet developed a comprehensive open-source toolkit to use such filtered VSVI in human perception research.

Considering this gap, we aim to develop a human perception survey toolkit based on the *findable, accessible, interoperable and reusable* (FAIR) principles (Wilkinson et al., 2016; Barton et al., 2022) that takes a *citizen science* approach, which means it enables and encourages non-scientists to collaborate with scientists and contribute towards the advancement of scientific research (Fraisl et al., 2022; Haklay, 2015). This paper describes the detailed process of creating our open-source VSVI-integrated software aligned with the FAIR principles for Research Software (Hong et al., 2022), and provides an example of using this toolkit for perception data collection as a form of participatory sensing (Haklay, 2015). The survey is a mobile-friendly web application that allows participants to rate SVI based on what they see in the image. Participants may be anyone with access to a computer or a smartphone connected to the Internet. Our software presents a lightweight 'game-like' user interface (Bakhanova et al., 2020) that requires no more and no less than one swipe or button press per image. The source code for our survey frontend, server backend, and the VSVI filtering and pre-processing scripts, are made available under the GNU General Public License v3.0 and may be found at our Spatial Data Science and GEO AI Lab[3] web site. Together these constitute a pipeline, suitable for civic/community science (Haklay, 2015), that can be assembled by any person with basic knowledge of deploying open source software on a hosted server environment supporting ReactJS, such as virtual servers readily available for low cost from numerous providers.

2. Background

Traditional approaches to capturing human perception of space include conducting direct observations such as field surveys and resident interviews (Lynch, 1960). For instance, 'walk-along interviews' may be used to capture

[1] www.un.org/sustainabledevelopment/sustainable-development-goals/

[2] creativecommons.org/licenses/by-sa/4.0/deed.en

[3] github.com/Spatial-Data-Science-and-GEO-AI-Lab/percept

the experience of space on the spot while interviewing individuals (Carpiano, 2009; Rzotkiewicz et al., 2018). Field audits and environmental scans are two more tools for evaluating streets. For example, Van Herzele and De Vries (2012) conducted a field audit to assess the visible greenness level on the streets of two urban neighbourhoods in Gent (Belgium). Harden et al. (2024) compared the effectiveness of SVI-based 'virtual audits' against more traditional in-person environmental scans. Although these approaches provide a more localized assessment of human perception and lived experience of spaces, they are limited due to very high resource and time requirements. Their method does not scale to a large number of places. Considering such limitations, researchers have innovated new ways to survey human perception using massive imagery sets and a 'crowdsourcing' approach (Bubalo et al., 2019). We take a closer look below at several recent works that are representative of the main methods used in such research.

Several studies have used a method of scoring imagery based on *relative (pairwise) comparisons*, such as Ye et al. (2019), Larkin et al. (2022), Streetscore (Naik et al., 2014) and most notably Place Pulse (Dubey et al., 2016; Salesses, 2012). The last was an extensive and wide-ranging effort to gather a crowdsourced SVI perception comparison dataset; in its second edition, the researchers collected 1.5 million comparisons for 110,998 Google SVI photos. As each participant worked through the web-based survey, they were shown two randomly-selected images at a time, side-by-side, and asked to choose which one of the two was 'better' (or equal) according to some selected criterion or 'perceptual attribute' such as *Safety* or *Beauty*. These relative comparisons were then translated into absolute scores and an overall image ranking per criterion. The crucial problem of Place Pulse is that to function correctly such comparison-to-score translation algorithms require an order of magnitude more comparisons than the survey was able to collect from human participants. Dubey et al. (2016, p. 6) states that they required '24 to 36 comparisons per image' but could only collect on average 3.35 comparisons per image. Therefore, they developed a customized machine learning algorithm to synthesize additional comparisons based on the collected ones. All this entails a substantial amount of work just to arrive at an absolute score ranking of the SVI in their dataset. Their root justification for using this relative comparison method comes from psychological research that studied humans performing tasks of making absolute identifications vs relative comparisons of simple stimuli such as 'sound tones' or 'line lengths' (Stewart et al., 2005). However, a significant limitation of that earlier psychological research is that it did not consider the evaluation of complex imagery such as SVI nor did it ask participants to rate imagery based on higher-level conceptual criteria such as 'safety' or 'walkability'.

In contrast, Twedt et al. (2016) used the direct *absolute scoring* approach, with a survey website where each image was shown one at a time and had to be rated on a scale of 0 to 100 using a slider adjusted with the mouse. Only 40 images were rated overall. There were approximately 300 participants, each was paid a small amount for their time via Amazon Mechanical Turk, and shown a preview of all the images before being asked to rate them individually. Pearson et al. (2024) operated similarly but scaled up to 10,727 paid participants, each of whom rated 33 images. They reported that the actual average completion time for the survey was 104 minutes, which was considerably higher than the 8 minute completion time they had expected from internal testing. Kruse et al. (2021) paid 210 Mechanical Turk workers to consider 'playability' on a five-point absolute scale and rated 3,011 images from 3 U.S. cities.

Yao et al. (2019) created a human/machine feedback-loop for speeding up manual absolute scoring work while simultaneously improving the accuracy of modelled perception ratings, on the same perceptual attributes as Place Pulse (e.g. Beauty, Safety, etc). This approach sped up the manual rating process to approximately 1,000 images per hour (3.6s per image) using a recommendation algorithm based on a machine learning model. However, there are several problems with both the method and the design of the software. Firstly, one of the biggest barriers to reuse of this survey method is that it is a software application that must be downloaded and installed locally on each participant's computer. This was not much of an obstacle for the authors because all of their participants were invited students or university staff, who were presumably comfortable with installing software from a fellow university member, but it is a major problem when trying to attract participation from the wider public. Secondly, the software's user interface is not user-friendly because of the amount of mouse-movement required for each image: the participant adjusts a slider to a value between 1 and 100 and then they must activate a submit button (although at least some of this interaction is also possible via hot-key bindings). Reducing this user-interaction effort appears to be one of the major underlying motivations of this work. However, this leads to the third problem: to speed up the rating process, the recommendation algorithm chooses a rating that is ready to be submitted by default, but in doing so it rewards participants who simply agree with the recommendation. There is no way to distinguish true human ratings from computer-generated ratings for which the human participant did not have strong enough feelings to justify the effort of moving the mouse to change the slider away from the default setting. Therefore the algorithm risks tainting its own training data too much. Fourthly, the authors noted a significant problem, which we see as being related to (but not quite the same as) the third: the given recommendation may have the unintended consequence of influencing the participant even when they do choose to adjust the slider. For example, suppose a participant would have rated an image with a score of 20 in Beauty, but then they saw the machine recommended a score of 40, so they decided to compromise on a score of 30. There is still an element of human input in this case, but it is biased by the recommendation algorithm.

All of these aforementioned works relied on either author-supplied photographs or commercial SVI, usually Google, but some used Tencent or Baidu. We instead set out to build reusable, open and FAIR software for assembling and operating street view perception surveys based on an open provider of VSVI, Mapillary.

3. Materials and methods

3.1. Mapillary street view imagery

Our method uses imagery from the Mapillary platform, which offers free-to-use SVI and VSVI in many cities around the world under the terms of a Creative Commons license. Some of this imagery was collected professionally with high-quality 360-degree panoramic cameras, especially in cities like Amsterdam (see Figure 1 for a sample of a cropped subimage of a panoramic photograph). Most of the available imagery is VSVI, of varying quality, but recent work by Zheng and Amemiya (2023) shows that such VSVI can be usefully filtered and employed in research as an alternative to higher-cost or non-free options.

We explicitly chose not to use proprietary data. For example, the standard terms of service[4] for Google SVI stresses that images are non-free and explicitly forbid the kind of usage that we need for research. In particular, in section 3.2.3 of the terms of service, they prohibit *scraping*, *pre-fetching*, *bulk downloading*, *storing*, and *resharing* Google SVI, all of which are necessary components of our survey method. Baidu and Tencent Maps also list several similar restrictions in their terms of service, and in any case, those platforms are mainly focused on providing imagery from China rather than worldwide.

While Google imagery is high-quality and widely available, we intend to keep within the open and FAIR principles; in particular, we are concerned that other researchers may not be able to obtain the authorization to use Google imagery within their projects (Helbich et al., 2024) and therefore we wish to show that there are alternatives that are free in the sense of being accessible and available ('*libre*') for research usage without overly-burdensome restrictions. The fact that Mapillary is free in the sense of monetary cost ('*gratis*') is also helpful. In contrast to Google, the Mapillary terms of service[5] only require proper attribution of imagery that is downloaded and reshared or integrated into applications.

3.2. Imagery downloading and processing

Our Python script, `mapillary_jpg_download.py` gathers imagery from Mapillary via their API (see Appendix C.1 for more details). The script only requires that you obtain your own free API key from the Mapillary developer portal, and that you provide a geographic bounding rectangle for your region of interest. It is also possible to obtain imagery from other sources, and inject it into the process at this point, should Mapillary be unsuitable or undesired for any reason, however the focus of this work is on Mapillary-provided imagery.

Once the tile data files and photographic imagery are downloaded, we must process a very large number of JPEG files, some of which are panoramic photographs projected into a wide image format, and others which are simply plain photographs. The two main tasks we need to accomplish are: (1) finding 'sensible subimages' in the panoramic imagery to crop out and save separately, and (2) weeding out any images that are too dark, too blurry or defective in some other way (such as only showing a wall or an undifferentiated block of greenery). These are both somewhat vague needs but thanks to recent advances in machine learning they can both be addressed using off-the-shelf software. In the case of panoramic photographs, we took subimages of a 4:3 ratio to show users images that fit within the app's available view-port for imagery, rather than distorted and overly large raw panoramic images.

Our definition of 'sensible subimages' was refined through trial and experiment, and we settled on the following: find the centers of roads in each of the panoramic images, and then for each road center crop a series of 4:3 images slightly to the left of center, on the center itself, and slightly to the right of center, as shown in Figure 2. This captures a wide variety of ways of looking at streets, including many views with a great deal of built environment and greenery, as well as views straight down the center of roads.

[4] `cloud.google.com/maps-platform/terms/`

[5] `www.mapillary.com/terms`

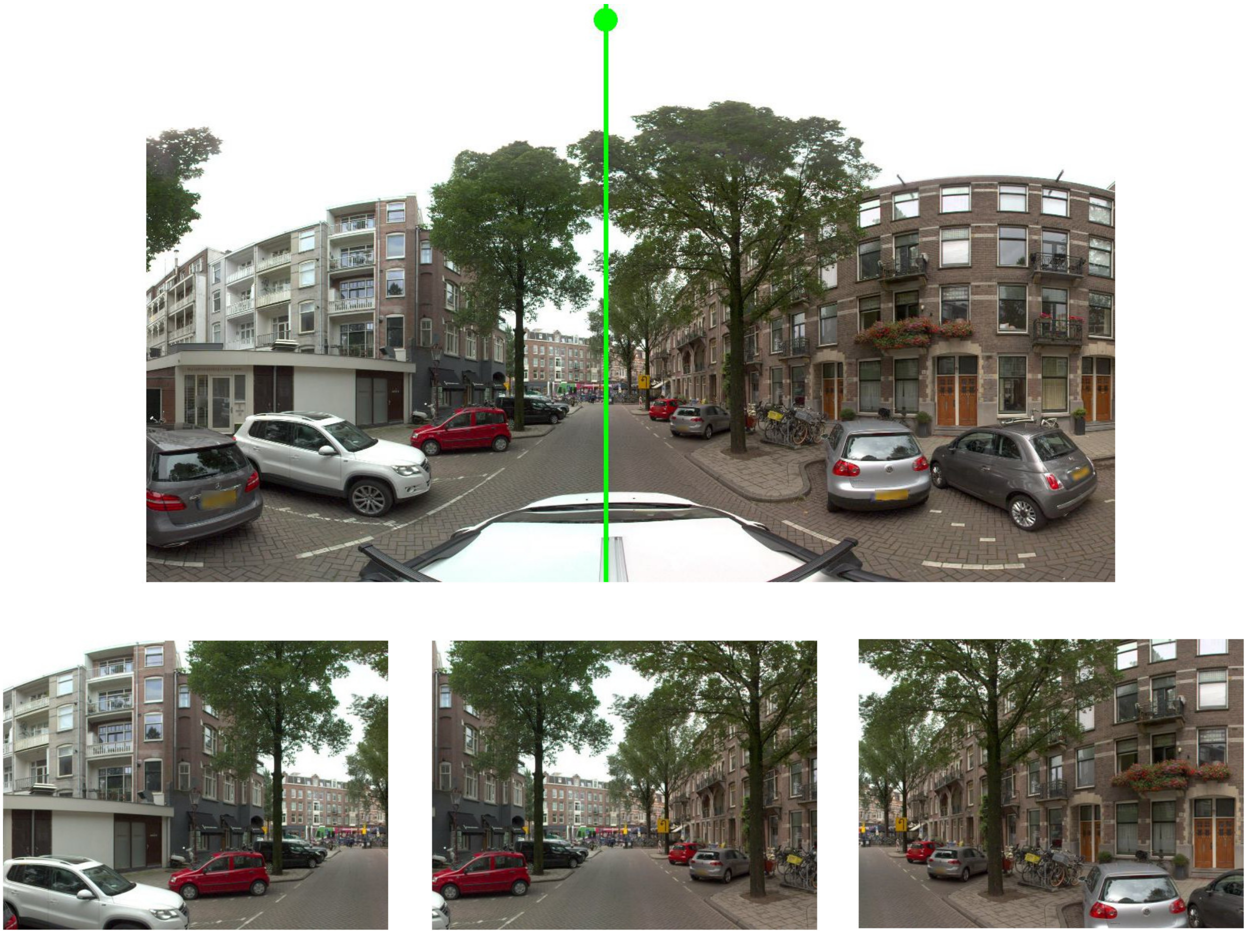


Figure 2: For each road center (indicated by the green line that is also marked with a circle) found in a panoramic image (top), we take three crops (bottom, from left to right): one slightly left of center, one directly facing the center, and one slightly right of center.

Finding road center-lines The panoramic imagery from Mapillary is generally normalized so that the leftmost edge of the image is where the compass would point north from the location where the SVI was taken. In theory, it should be possible to reconstruct the position of streets within panoramic imagery based on street map vector data (such as OpenStreetMap). However, in practice, we found some discrepancies when examining SVI samples: the due north direction was not always the leftmost edge. In addition, the arrangement of streets in reality can be considerably more complex than shown in the simplified model of a vector-based street map. By finding road center lines using a computer vision algorithm, our software is robust in the case of imagery where compass angle is wrong or simply not known. We find the center of roads in imagery by first labeling all of the pixels as either 'road' or 'non-road' using semantic segmentation (Thisanke et al., 2023), and then seeking the 'peaks' of the road pixel distribution horizontally across the image, with some additional code to handle common distortions and edge cases. See Appendix C.4 for more details.

3.3. Filtering by image quality

Most of the volunteered photographs that we encountered on Mapillary are non-panoramic and can range from very good quality to quite poor. Inspired by the filtering ideas of Zheng and Amemiya (2023), we calculate contrast and 'tone-mapping score' (Stefanescu, 2021) using the Python Scikit-image library (Van der Walt et al., 2014), and apply a certain threshold on a formula composing both values (see Appendix C.5), to select acceptable images. We also applied the road center-line finding algorithm described above, but in this case, only to determine if there is a road or not within the image, and whether that road center-line lies closer to the center of the image than to the edge. Images that pass these tests are then cropped to fit a 4:3 ratio so they fit into the image view-port of our web app.

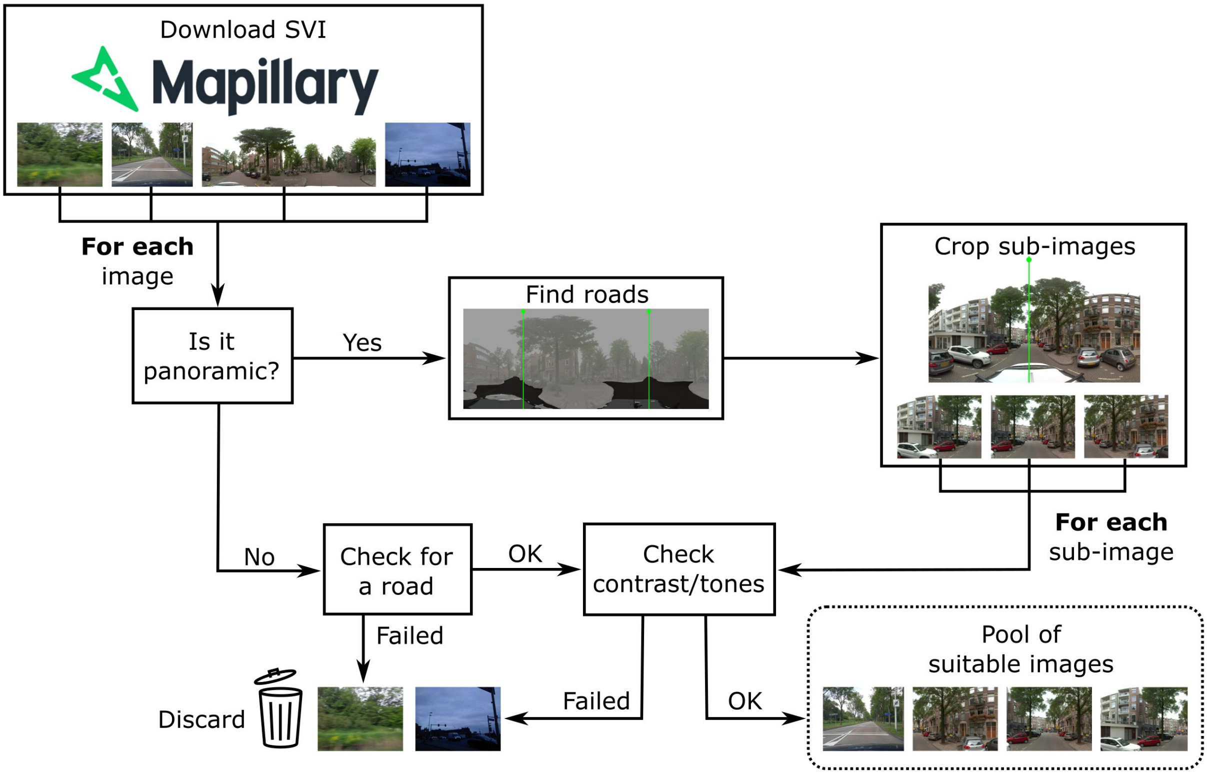


Figure 3: Flow diagram of the preparation and processing of SVI

It is possible for so-called 'bad' images to slip through, ones that have poor photo quality or some form of obstruction that renders them ill-suited for evaluation; however after experimentation we found that such instances are rare, and they can be reported when found by a participant. To some extent there is also a subjective feeling of 'badness', which cannot be controlled. For example, some people prefer images with blue skies and do not like rating SVI taken on cloudy days, whereas other people do not mind such images. Since the survey is purely voluntary, we permit people to skip images that they feel they cannot rate, for any reason, as to not discourage participation.

An overview of all the steps of downloading and processing of Mapillary VSVI are shown in the diagram within Figure 3. After these steps, the imagery is ready to be used in the perception-gathering survey, or it may be taken and used for other purposes in compliance with Mapillary terms of service.

3.4. Survey frontend

The perception-gathering survey is written as a mobile web app in the ReactJS framework. First of all, before the main perception-gathering survey takes place, there is a short socio-demographic survey requesting a few personal details: age, education level, gender, approximate income, home postal code, country of residence and consent for data collection (see Section A.2 for more details). Only age and consent are required fields. We comply with the EU General Data Protection Regulations[6] and require participants' explicit consent, which may be withdrawn at any time. Furthermore, we do not use data from participants who are under 18.

The main perception-gathering survey then proceeds: it shows participants one image at a time along with one of five possible categories: walkability, bikeability, pleasantness, greenness or safety (see Appendix A.1 for more details). Participants may rate each image according to the given category by pressing one of five rating buttons, or swiping the image towards one of the rating buttons. The buttons are arranged along a Likert-type scale: awful, bad, neutral, good, and great (internally numbered from 1 to 5). Participants rate five images according to a given category and then the app chooses a new category at random, until 20 images have been rated in each category. There are also options to skip

[6] commission.europa.eu/law/law-topic/data-protection_en

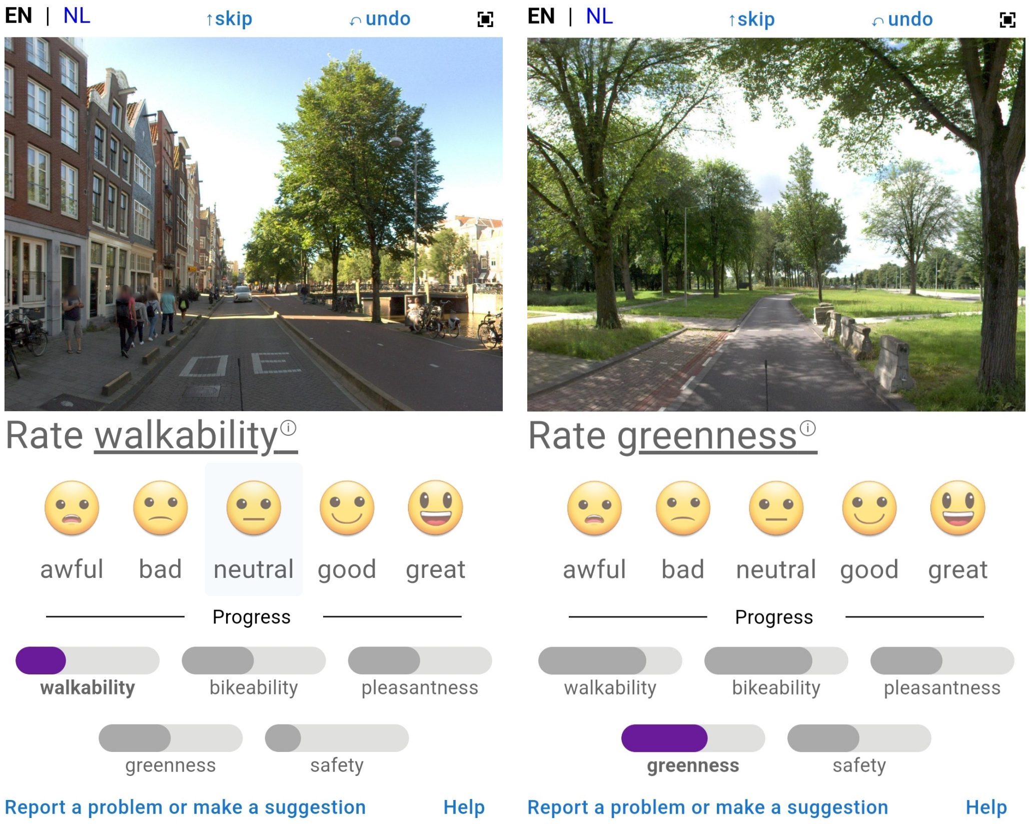


Figure 4: Two examples of mobile screenshots

the current image, undo the last rating (see Appendix B.2 for more undo details), go full screen, change language, get more help, and report an issue. The mobile web app (Figure 4) is designed to fit seamlessly on a mobile device so that users can pull up their smartphone to quickly do some ratings whenever they desire.

The frontend runs entirely on the user's browser and communicates with the backend server via a public API (see Appendix B.1 for more API details). Image URLs are fetched from the backend, and then downloaded from the image-hosting server as needed. Ratings and undo commands are sent asynchronously to the backend while the interface updates.

The user interface was pilot-tested with Utrecht University students and staff. We made several changes as a result, including: the positioning of the SVI, the text of the category descriptions and the wording of the socio-demographic survey items. We also decided to limit participants to 20 ratings per category (100 ratings total) and show progress bars in each category to give people a sense that they were working towards a definite ending.

3.5. Survey backend

The backend is an independent ExpressJS-based server with a well-defined public API for tasks such as starting new sessions and submitting surveys and ratings. The backend is a separate module from the frontend; it is possible that a different frontend (or even raw requests) could interface with the backend, if so desired. The public API (see Section B.1) is treated as a potential entry point for input from any possible source, including malicious ones. Data is stored securely on a private server using PostgreSQL and PostGIS for geographic data processing. The overall arrangement of frontend, backend and client (in this case, depicted as a smartphone user) is shown in Figure 5.

3.6. Computational Resources

The survey server that we used for our case study is a Linux virtual machine hosted on an Intel Xeon™ 5120 CPU with 4 cores running at 2.2GHz, having 16GB RAM available and approximately 100GB of disk space used for data

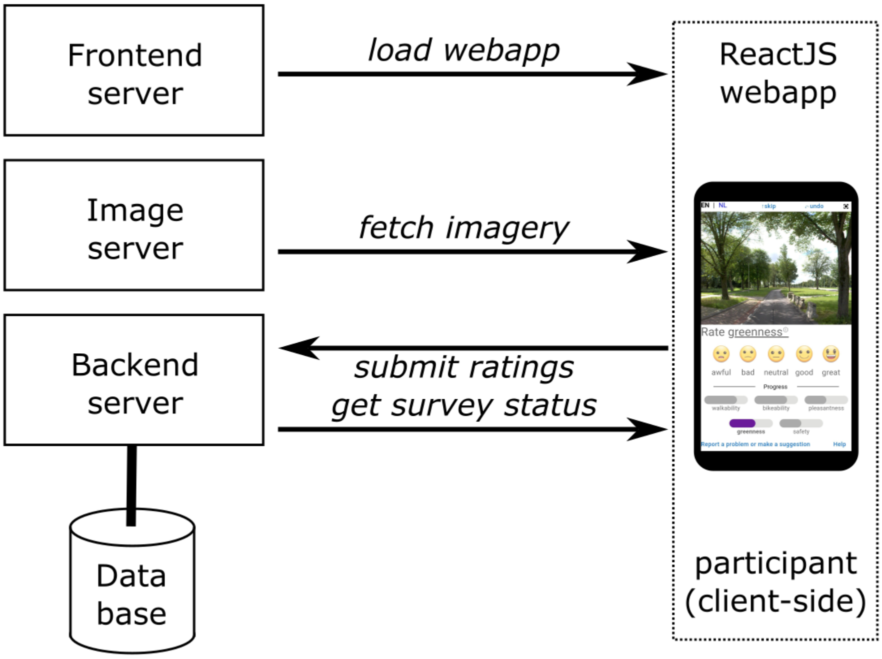


Figure 5: A schematic showing server-side (left) / client-side (right) and their interactions.

and imagery. Reasonably comparable virtual servers can be hired (as of this writing) for around €10-20 per month. We had no more than 13 people simultaneously participating in our survey and our server had more than sufficient resources to support that load. In terms of hardware and software flexibility, it should be possible to run our open source software on any operating system that supports ReactJS v18.4 and the PostgreSQL v13 database, however we have only tested (and therefore recommend) the following: Linux (CentOS 7) with Apache2 v2.4.6 (configured as a proxy), PostgreSQL v13.14 with PostGIS v3.3.3, and Node v18.4.2 with React v18.2.0 and Express v4.18.2. Newer versions of dependencies should work fine, especially regarding Apache2 and PostgreSQL (which on CentOS are substantially older but stable versions maintained by the vendor). Code changes may be necessary for updated Node and React libraries and that will be managed on our GitHub repository. Preprocessing operations were handled by Python 3.6 scripts under Linux but should work fine with newer versions. We recommend having at least approximately 2 TB of disk space available for comfortably managing to download and filter VSVI on a region comparable to the one we describe in Section 4.1; this larger set of raw VSVI can later be deleted to save space if desired.

4. Results

4.1. Survey set-up for the Amsterdam case study

We retrieved imagery from points found within the bounding box described by longitudes 4.7149 and 5.1220, and latitudes 52.2818 and 52.4284 (WGS 84). This encompasses the city of Amsterdam and some outlying areas. The total number of images found (panoramic and otherwise) in this bounding box was well over 700,000. The panoramic images available from Mapillary for Amsterdam are almost entirely high-resolution and professional-quality, taken from a 360-degree panoramic camera mounted on a vehicle or backpack. Using our method, each panoramic image could potentially be used to derive up to nine high-quality subimages. With so many possibilities, we heavily filtered the amount of possible SVI first by applying the pipeline as described in Section 3.3, and then selecting only images from locations closest to a fixed geographic grid of points covering the whole region but spaced approximately 20m apart from each other. This still left us with too many images, so we then down-sampled from these eligible images randomly until the number of images for the survey was a bit under 20,000, which we felt was a sufficient number for our initial run, and well-distributed across the region of interest.

4.2. Collected perception

From April 2023 to February 2024, we received 22,637 ratings across 19,750 images from up to 331 participants. We recruited participants using a wide variety of channels including but not limited to: social media, institutional mailing lists, and classrooms. We asked several socio-demographic questions of participants before they began the main part of the survey. Of those who responded, 40.5% (134) identified as women, 0.9% (3) as non-binary, 52.3% (173) as men, and the remainder declined to answer. Educational attainment amongst participants was fairly high, probably due to heavy promotion of the survey within the academic environment; 52.9% (175) people completed postgraduate education, 39.9% (132) were at the tertiary level (in the Netherlands this includes university as well as higher vocational education), 3.0% (10) had completed secondary education (high school), 0.3% (1) claimed only a primary education level and the remainder did not specify. The age spread was largely middle-aged or younger, with 48.6% (161) participants of ages from 18 to 27, 32.6% (108) aged 28 to 37, 11.2% (37) aged 38 to 47, 5.1% (17) aged 48 to 57, and the remainder older than 57. The average age of participants was about 30 (standard deviation 10). Approximately 70.7% (234) of the participants reported being based in the Netherlands, approximately 8.8% (29) reported living in another European country, a further 7.6% (25) in a non-European country, and the rest did not indicate a country or were indecipherable. About 39.0% (129) claimed a monthly gross income of less than €2,000, probably reflecting a high percentage of students, followed by about 24.2% (80) with a reported income between €2,000 and €4,000, and about 12.4% (41) claimed to have a higher income than €4,000 per month; the remainder did not specify.

Of our 331 participants, 59.2% (196) submitted at least 50 ratings, and 50.8% (168) finished the full 100. The median survey completion time was 6.3 minutes and 93.4% (309) of participants took 30 minutes or less from start to end. We observed participants swiping or clicking with ease through our smartphone-friendly user interface, and we later confirmed this using timestamps in our database: the median interval between ratings was 3.7 seconds and 89.7% (20,310) of the submitted ratings occurred within 10 seconds of the previous rating.

4.3. Spatial distribution of perception

As an illustrative example, Figure 6 shows the collected walkability geocoded perception data, averaged into hexagonal bins each measuring approximately 650m wide west/east and 600m north/south. These fully-anonymized data points for this map are available on our web site. We find it encouraging that this overview of the data aligns with the common sense intuition that worse walkability perceptions occur more often on the outskirts of the city and better perceptions should be more frequent closer to the center of the city. We note that many of the better ratings that are found in the outskirts tend to be linked to photos of residential neighbourhoods or parks, and the worse ratings with industrial areas or high-speed roads, which are more common outside the center.

5. Discussion

5.1. Main findings

The goal of this work is to provide an open-source end-to-end pipeline for building and conducting street perception surveys. This enables citizen science-based (Haklay, 2015) perception research in two ways: firstly, the survey itself is a form of participatory sensing through an easy-to-use smartphone-friendly web app. Secondly, by keeping to the FAIR principles (Wilkinson et al., 2016; Barton et al., 2022) for data and software, we promote open science and reproducibility. Our choice to rely on FAIR data led us to use preparation methods inspired by Zheng and Amemiya (2023) for cleaning and selecting VSVI rather than obtaining potentially cleaner SVI from proprietary platforms that have overly restrictive terms of service.

Our method of scoring is easier to use than that of Place Pulse (Dubey et al., 2016; Salesses, 2012) or other relative comparison studies (Naik et al., 2014; Ye et al., 2019; Larkin et al., 2022). Like Kruse et al. (2021), we collect ratings along an absolute five-point scale, whereas much of the challenge in projects such as Place Pulse was in reconstructing a ranking from a large set of relative comparisons; this is a basic philosophical difference in approach. The advantage of using absolute scores is that the ratings are always linearly ordered. We also argue that with high-level and high-complexity concepts such as 'walkability' and 'safety' the use of relative comparisons is not necessarily helpful, since there are so many aspects of the imagery to consider simultaneously; past psychological research to justify the use of relative comparisons (Stewart et al., 2005) considered only low-complexity concepts such as 'line lengths' or 'sound tones'. Instead of having a complex process of translating relative comparisons into absolute scores, we keep things simple in our system by showing only one image and enabling participants to give a quick first-impression one-click/swipe answer from a selection of five possible responses. This user interface helps participants quickly and

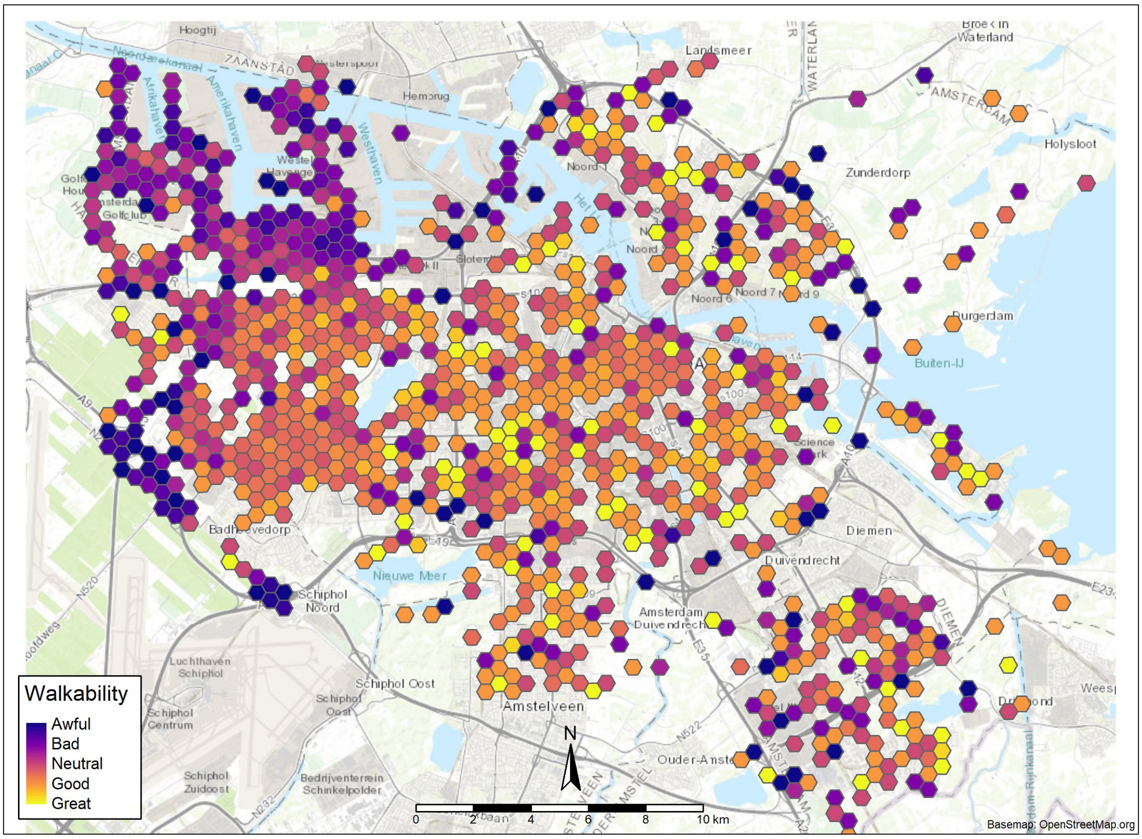


Figure 6: Collected walkability ratings in and around Amsterdam, averaged into hexagonal bins

smoothly look through many more images than interfaces with a rating slider (Twedt et al., 2016; Larkin et al., 2022) and avoids any biasing problems created by feeding back preliminary machine learning model results into the user interface (Yao et al., 2019), while still allowing participants to able to enjoy rapid progression through the survey.

We were able to easily attract public participants and give them the flexibility to respond at their convenience because our survey is a smartphone-friendly web app rather than a downloaded software application. We did not pay participants (Twedt et al., 2016; Pearson et al., 2024; Larkin et al., 2022; Kruse et al., 2021) but rather relied on the attractiveness of the web app as a fun ‘game’ that can be played during downtime (e.g. waiting for the bus).

In the future, we will expand our toolkit by building a framework for generating models of the data collected by our street perception surveys. The state-of-the-art in image classification has moved on considerably in the past five years. Within the past year alone, researchers have published large pre-trained models such as StreetCLIP (Haas et al., 2023), which is designed to perform zero-shot image classification and has been specialised by training on a geo-tagged SVI dataset containing 1.1 million images. For example, as input to StreetCLIP one could provide a photograph taken from a city street and a set of three descriptive captions such as ‘An urban street’, ‘A suburban road’ and ‘A rural road’. The output of the model would be a probability distribution indicating which one of the input descriptions is the most likely to be suitable for the given image. StreetCLIP is based on the CLIP text and image association models (Radford et al., 2021). We can take StreetCLIP even further by fine-tuning it on our collected images and ratings by transforming our categories and five-point numerical rating scale into textual image captions using a template such as ‘An image of a street with a [rating text] [category] rating’. For example a walkability rating of 5 becomes ‘An image of a street with a very good walkability rating’, and a safety rating of 2 becomes ‘An image of a street with a bad safety rating’. The resulting image and text models will be able to take an input of an image and a selection of five templated sentences, and then output a probability distribution indicating which of those the most likely sentence to apply to the given image. In this manner, we can extrapolate the perception data by starting from the collected ratings for tens of thousands of images and then obtain modelled ratings for millions of images. One disadvantage of StreetCLIP is that the published model is based on Google SVI, therefore one option we must consider is retraining it based on open or VSVI instead, in addition to our collected dataset.

5.2. Limitations

Mapillary[7] has more than 2 billion images from 190 countries, and especially good coverage in countries like the Netherlands, including almost full panoramic coverage of Amsterdam-area streets. However, although Mapillary has imagery from every continent, the most abundant SVI coverage comes from Europe and North America (Ma et al., 2019). However, commercial SVI is no panacea either: Kim and Jang (2023) analyzed Google SVI coverage of walk commute trajectories in small- to medium-sized cities in the United States and found gaps in nearly half of the routes they investigated; they suggest that researchers consider Mapillary to bolster coverage.

The five-point scale is simple for participants but as a result sometimes is too coarse-grained to express the true rating that the participant feels. Similarly, descriptions of categories try to give unambiguous criteria for people who need some guidance with their consideration of images, however with some categories there is an unavoidable tension between different interpretations. For example, with ‘safety’ many of the survey testers found themselves thinking of a ‘road safety’ interpretation rather than a ‘personal safety’ interpretation, although we tailored the description to fit the latter. However, we do not wish to dictate the responses from participants, in the end we are trying to collect what people are already thinking rather than trying to teach them something, and it is reasonable to consider road safety as a component of personal safety.

The absolute scoring system avoids the complexity of the relative comparison system used by Place Pulse (Dubey et al., 2016; Salesses, 2012) but can result in situations where participants might change their minds about images and wish to go back and redo the ratings. We purposefully limited ‘undo’ functionality to a single previous rating to prevent participants from undoing large numbers of ratings. This protects the integrity of the system against unexpected mass cancellation of ratings. Our system collects ‘quick impressions’ with the idea being that participants spend no more than a few seconds on each image; revision or reconsideration of ratings would work against that goal, even if the participant later thinks differently.

Our survey relies on Internet access and works best on a modern smartphone although it can be completed on a regular computer with a reasonably up-to-date (within the last half-dozen years) web browser as well. Not everyone has Internet access, although recently published survey work from the Pew Research Center (Poushter et al., 2024) found that in most countries surveyed approximately 8 out of 10 adults, or more, use the Internet.

6. Conclusion

We offer to the research community an open-source toolkit for downloading, processing and filtering VSVI from a given geographic region, and deploying a mobile-friendly human perception survey web app on the resulting images. In the spirit of citizen science and the FAIR research principles, anyone may easily replicate, modify and deploy this toolkit on any location of interest. We integrated it with the Mapillary platform because they provide open access to VSVI from many cities around the world. This toolkit will be used to build deep learning models with the collected data and make predictions about human perception of large amounts of SVI over a wider area. Ultimately, we hope to see such perception data used in research that can help guide future planning and development choices, and therefore improve the quality of life for many people in urban areas.

Software & data availability

- Name of software: Human perception and volunteered street view imagery project (percept)
- Developer: Matthew Danish, m.r.danish@uu.nl
- Source code / data: `github.com/Spatial-Data-Science-and-GEO-AI-Lab/percept`
- Date first made available: February 2024
- Hardware required: (server) Internet-connected server or virtual server; (client) Smartphone or desktop.
- Software required: (server) Linux, Apache2, PostgreSQL, Node.js and Express.js; (client) Web browser.
- License and cost: GNU General Public License 3.0; there is no cost.
- Programming languages: JavaScript, HTML/CSS, Python

[7] `mapillary.com/about`, as of February 2024

The available data has been processed to ensure it contains absolutely no personally-identifiable information and to prevent reconstruction of any such information, a problem which might occur in combination with other data sources.

Declaration of competing interest

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

Acknowledgements

This work received funding (BMD 3.1fb220215) from the Faculty of Geosciences, Utrecht University, as well as from the FAIR Research IT program at Utrecht University. Neither of the funders had any role concerning the study design, data collection and analysis, interpretation, or dissemination. Maarten Zelymans Van Emmichoven helped with early conceptualization, and the Nederlands translation of the app.

CRediT authorship contribution statement

Matthew Danish: Conceptualization, Writing, Software, Methodology, Visualization. **S.M. Labib:** Conceptualization, Writing, Funding acquisition, Methodology. **Britta Ricker:** Conceptualization, Writing, Funding acquisition, Methodology, Visualization. **Marco Helbich:** Conceptualization, Writing, Funding acquisition, Project administration, Methodology.

References

Bakhanova, E., Garcia, J.A., Raffe, W.L., Voinov, A., 2020. Targeting social learning and engagement: what serious games and gamification can offer to participatory modeling. Environmental Modelling & Software 134, 104846.

Barton, C.M., Ames, D., Chen, M., Frank, K., Jagers, H.A., Lee, A., Reis, S., Swantek, L., 2022. Making modeling and software FAIR. Environmental Modelling & Software 156.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: a review. Landscape and Urban Planning 215, 104217. doi:doi.org/10.1016/j.landurbplan.2021.104217.

Bradski, G., 2000. The OpenCV Library. Dr. Dobb's Journal of Software Tools .

Bubalo, M., van Zanten, B.T., Verburg, P.H., 2019. Crowdsourcing geo-information on landscape perceptions and preferences: a review. Landscape and Urban Planning 184, 101–111.

Carpiano, R.M., 2009. Come take a walk with me: the "Go-Along" interview as a novel method for studying the implications of place for health and well-being. Health & Place 15, 263–272.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1290–1299.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The Cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223.

Dissart, J.C., Deller, S.C., 2000. Quality of life in the planning literature. Journal of Planning Literature 15, 135–161. doi:10.1177/08854120022092962.

Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A., 2016. Deep learning the city: quantifying urban perception at a global scale, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, NL, October 11–14, 2016, Proceedings, Part I 14, Springer. pp. 196–212.

Fraisl, D., Hager, G., Bedessem, B., Gold, M., Hsing, P.Y., Danielsen, F., Hitchcock, C.B., Hulbert, J.M., Piera, J., Spiers, H., et al., 2022. Citizen science in environmental and ecological sciences. Nature Reviews Methods Primers 2, 64.

Haas, L., Alberti, S., Skreta, M., 2023. Learning generalized zero-shot learners for open-domain image geolocalization. arXiv:2302.00275.

Haklay, M., 2015. Citizen science and policy: a European perspective. Technical Report. Woodrow Wilson International Center for Scholars. Washington, DC, USA.

Harden, S.R., Schuurman, N., Larson, H., Walker, B.B., 2024. The utility of street view imagery in environmental audits for runnability. Applied Geography 162, 103167.

Helbich, M., Danish, M., Labib, S., Ricker, B., 2024. To use or not to use proprietary street view images in (health and place) research? That is the question. arXiv:2402.11504.

Hong, N.P.C., Katz, D.S., Barker, M., Lamprecht, A.L., Martinez, C., Psomopoulos, F.E., Harrow, J., Castro, L.J., Gruenpeter, M., Martinez, P.A., et al., 2022. FAIR principles for research software (FAIR4RS principles). Research Data Alliance .

Illingworth, J., Kittler, J., 1988. A survey of the Hough transform. Computer Vision, Graphics, and Image Processing 44, 87–116. doi:doi.org/10.1016/S0734-189X(88)80033-1.

K, P.B.R., Anand, A., Soni, A., 2022. Road lane detection using computer vision. International Journal for Research in Applied Science & Engineering Technology (IJRASET) 10, 1993–2002. doi:doi.org/10.22214/ijraset.2022.45634.

Kim, J., Jang, K.M., 2023. An examination of the spatial coverage and temporal variability of Google Street View (GSV) images in small-and medium-sized cities: a people-based approach. Computers, Environment and Urban Systems 102, 101956.

Kothencz, G., Blaschke, T., 2017. Urban parks: visitors' perceptions versus spatial indicators. Land use policy 64, 233–244.

Kruse, J., Kang, Y., Liu, Y.N., Zhang, F., Gao, S., 2021. Places for play: understanding human perception of playability in cities using street view images and deep learning. Computers, Environment and Urban Systems 90, 101693.

Labib, S., Lindley, S., Huck, J.J., 2020. Spatial dimensions of the influence of urban green-blue spaces on human health: a systematic review. Environmental research 180, 108869.

Larkin, A., Krishna, A., Chen, L., Amram, O., Avery, A.R., Duncan, G.E., Hystad, P., 2022. Measuring and modelling perceptions of the built environment for epidemiological research using crowd-sourcing and image-based deep learning models. Journal of Exposure Science & Environmental Epidemiology 32, 892–899.

Lotfi, S., Koohsari, M.J., 2009. Analyzing accessibility dimension of urban quality of life: where urban designers face duality between subjective and objective reading of place. Social Indicators Research 94, 417–435.

Lynch, K., 1960. The image of the city. The MIT Press, Cambridge, MA, USA.

Ma, D., Fan, H., Li, W., Ding, X., 2019. The state of Mapillary: an exploratory analysis. ISPRS International Journal of Geo-Information 9, 10.

McCrea, R., Shyy, T.K., Stimson, R., 2006. What is the strength of the link between objective and subjective indicators of urban quality of life? Applied research in quality of life 1, 79–96.

Naik, N., Philipoom, J., Raskar, R., Hidalgo, C., 2014. Streetscore - predicting the perceived safety of one million streetscapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Pearson, A.L., Lin, Z., Shortridge, A., 2024. Beauty is in the eye of the beholder: moderate ability to predict perceptions of beauty, restorativeness, and naturalness, in imagery among 10,727 participants. Landscape and Urban Planning 243, 104958.

Poushter, J., Gubbala, S., Austin, S., 2024. 8 charts on technology use around the world. URL: www.pewresearch.org/short-reads/2024/02/05/8-charts-on-technology-use-around-the-world/.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision, in: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

Rzotkiewicz, A., Pearson, A.L., Dougherty, B.V., Shortridge, A., Wilson, N., 2018. Systematic review of the use of Google Street View in health research: Major themes, strengths, weaknesses and possibilities for future research. Health & place 52, 240–246.

Salesses, M.P., 2012. Place Pulse: Measuring the collaborative image of the city. Ph.D. thesis. Massachusetts Institute of Technology.

Stefanescu, M., 2021. Measuring and enhancing image quality attributes. URL: https://towardsdatascience.com/measuring-enhancing-image-quality-attributes-234b0f250e10.

Stewart, N., Brown, G.D., Chater, N., 2005. Absolute identification by relative judgment. Psychological review 112, 881.

Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., Herath, D., 2023. Semantic segmentation using Vision Transformers: a survey. Engineering Applications of Artificial Intelligence 126, 106669. doi:doi.org/10.1016/j.engappai.2023.106669.

Twedt, E., Rainey, R.M., Proffitt, D.R., 2016. Designed natural spaces: informal gardens are perceived to be more restorative than formal gardens. Frontiers in Psychology 7. doi:10.3389/fpsyg.2016.00088.

Van Herzele, A., De Vries, S., 2012. Linking green space to health: a comparative study of two urban neighbourhoods in Ghent, Belgium. Population and Environment 34, 171–193.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17, 261–272. doi:10.1038/s41592-019-0686-2.

Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. Scikit-image: image processing in Python. PeerJ 2, e453.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. Comment: the FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, 1–9. doi:10.1038/sdata.2016.18.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M., 2020. HuggingFace's Transformers: state-of-the-art natural language processing. arXiv:1910.03771.

Yao, Y., Liang, Z., Yuan, Z., Liu, P., Bie, Y., Zhang, J., Wang, R., Wang, J., Guan, Q., 2019. A human-machine adversarial scoring framework for urban perception assessment using street-view images. International Journal of Geographical Information Science 33, 2363–2384.

Ye, Y., Zeng, W., Shen, Q., Zhang, X., Lu, Y., 2019. The visual quality of streets: a human-centred continuous measurement based on machine learning algorithms and street view images. Environment and Planning B: Urban Analytics and City Science 46, 1439–1457.

Zheng, X., Amemiya, M., 2023. Method for applying crowdsourced street-level imagery data to evaluate street-level greenness. ISPRS International Journal of Geo-Information 12, 108.

Appendices

A. Survey details

A.1. Categories

Below are the categories used to gather ratings in the survey. Each participant was asked to rate 20 images in each category. The app presented five images consecutively under one category, and then the app randomly switched to another unexhausted category. For each category below is the corresponding category description, which was shown to each participant the first time they encountered the category. The same text was also subsequently available under a tooltip, for reference. The purpose of the category descriptions was not to be prescriptive but rather to help alleviate concerns about ambiguity or inspire participants who were uncertain about how to respond.

- Walkability – *Does this place look like an easy and safe place for people to travel on foot or using a walking-equivalent mobility aid (e.g. wheelchair)? This might include factors such as the quality of sidewalks, pedestrian crossings, street connectivity, and access to public amenities. Walkable communities encourage people to walk or use other non-motorized modes of transportation.*

- Bikeability – *Does this place look accessible, attractive, safe and convenient for cycling as a mode of general-purpose transportation, or cycling-equivalent mobility aid (e.g. mobility scooter)? This might include factors such as cycle lanes, tracks, and parking, as well as the overall design of streets, junctions and any visible surroundings.*

- Pleasantness – *Does this place look enjoyable or pleasing to the senses or emotions? This might include factors such as the aesthetics of the surroundings, the quality of the air and lighting, the soundscape, and the presence of other people or natural elements.*

- Greenness – *Rate the apparent amount of vegetation and greenery in a given environment. This encompasses the presence of trees, shrubs, plants, and other natural elements.*

- Safety – *Does this place look like you would feel protected from harm or danger, in terms of personal safety and security? Do you believe that it is likely that you would feel safe here at all times of day or night? This might include the presence of elements such as lighting, good maintenance, presence of other people and natural surveillance.*

A.2. Socio-demographic information

- Age
- Gender
- Estimated monthly income
- Level of education
- Home postal code and country of residence
- Data usage consent

We gather several pieces of socio-demographic information to make comparisons between responses from users of different backgrounds. We strove to keep the number of demographic questions limited to five because we did not want to create an unnecessarily high barrier to entry. For example, it would be helpful to know each participant's country of birth and upbringing, as well as country of residence, but we felt that would be too confusing. Of the information collected, only age and data consent are required for participation in the survey. Age is required because we want the users to self-certify that they are 18 or older, to avoid complications with obtaining data usage consent from minors.

The other questions are optional and relatively free-form, with the understanding that we may not be able to interpret all answers given. 'Level of education' is the most structured question, with four possible options ('Primary', 'Secondary', 'Tertiary' and 'Postgraduate'). This is a compromise between legibility and specificity because, for example, in the Netherlands it is quite common to break down education level into categories like MBO, HBO, and VWO, but these abbreviations have no meaning to people who are not familiar with the Dutch system.

B. Backend details

B.1. Public API v1

The backend server defines a URL for each API function, of the form `/api/v1/<function>` where <*function*> is one of the following, along with corresponding form parameters:

- `newperson` (`age`, `monthly_gross_income`, `education`, `gender`, `country`, `postcode`, `consent`)
  - Creates a new participant based on the given socio-demographic inputs (only `age` and `consent` are required).
  - Returns as a JSON dict `session_id` and `cookie_hash`, the latter of which is intended to be stored as a cookie in the participant's browser if they consent to the study.
- `getsession` (`session_id` or `cookie_hash`; either can be used, depending upon what is known)
  - Finds the corresponding `session_id` for a given `cookie_hash`, or vice versa.
  - Returns as a JSON dict `session_id` and `cookie_hash` both fully filled out.
- `fetch` (`session_id`)
  - Select an image from the database that has not yet been rated by the current participant.
  - Returns as a JSON dict `cityname`, `url` and `image_id`
- `new` (`session_id`, `cookie_hash`, `image_id`, `category_id`, `rating`)
  - Creates a new rating data point for the current session.
  - Returns the same result as `countratingsbycategory`.
- `undo` (`session_id`, `cookie_hash`)
  - If permitted (see Appendix B.2) then undo the most recent rating by the participant.
  - Returns the same result as `countratingsbycategory`.
- `countratingsbycategory` (`session_id`)
  - Returns a JSON dict with a single element, `category_counts`, which in turn contains a JSON dict keyed by category ID with information about how many images have been rated by the current participant in each corresponding category. For example, `result['category_counts'][1]` gives the number of ratings that have been submitted for the category with ID 1.

B.2. Undo protocol

The backend enforces a particular undo protocol: only the single most recent rating can be undone. This ensures that the use of undo is limited to only correcting a simple mistake and does not result in large-scale deletion of data from the database. Skipped images are not reported to the backend at all; undo of skips is possible and it is handled entirely in the frontend, in a transparent manner so that the participant cannot tell the difference between undoing a skip or undoing a rating.

C. Processing and filtering VSVI details

For full up-to-date usage information please see the Spatial Data Science and Geo AI Lab[8] web site.

C.1. Downloading imagery from Mapillary

`mapillary_jpg_download.py`

This script takes a Mapillary API key[9] and a bounding box (west, south, east, north) and conducts a lengthy but robust and restartable procedure to methodologically find and download each tile data file, cache it, and then download all of the eligible SVI that is found within the tile (and within the bounding box).

To give more detail: each tile data file is a set of GeoJSON features from within a certain pre-defined rectangular area. In our case, we are interested only in features corresponding to SVI. Therefore the tile files are lists of images (photographs) with the following pieces of information for each one: a unique image identifier, sequence identifier

[8] github.com/Spatial-Data-Science-and-GEO-AI-Lab/percept

[9] www.mapillary.com/developer

corresponding to a series of photographs taken in a row (often while driving down a street), the compass angle at which this photograph was taken, geographic latitude and longitude coordinates, the time it was taken, and a Boolean value indicating whether or not it is a panoramic photograph.

With the image identifier, we are able to use the Mapillary API to obtain the precise URL of the original photographic image, and then download it. The photographs are organised by sequence identifier (for later reference) and stored in filenames corresponding to the image identifier (which is unique). In this way, we have access to all of the available information about each photograph: the imagery is stored in a directory structure organised by sequence identifier, and the unique image identifying number can be used to look up all of the meta-information in the tile data files.

Should any part of the downloading process fail, the script automatically backs up and restarts the download process, with exponential back-off up to a limited number of retries. After that point, if the download still fails, the image identifier can be (optionally) saved to a file containing a list of failed identifiers. Later, the process can be restarted using the cached tile data files, and the script can be directed to focus on the failed image identifiers (or any image identifiers the user chooses). In our experience, the downloading process does fail from time to time, and therefore this functionality was very valuable, as the full download process can take days depending on how large the required region is.

C.2. Semantic segmentation of imagery

`torch_segm_images.py`

This script is designed to work with very large directories or lists of filenames corresponding to SVI. It uses the PyTorch library and by default the `facebook/mask2former-swin-large-cityscapes-semantic` model (Cheng et al., 2022) for image segmentation. The result is a matrix with values corresponding to the meaning of each pixel in the input image. These matrices are then stored in compressed numpy array files (.npz) alongside each image. This step takes a substantial amount of time but thanks to the PyTorch library it can be significantly sped-up with the assistance of a GPU.

C.3. Cropping and filtering segmented imagery

`torch_process_segm.py`

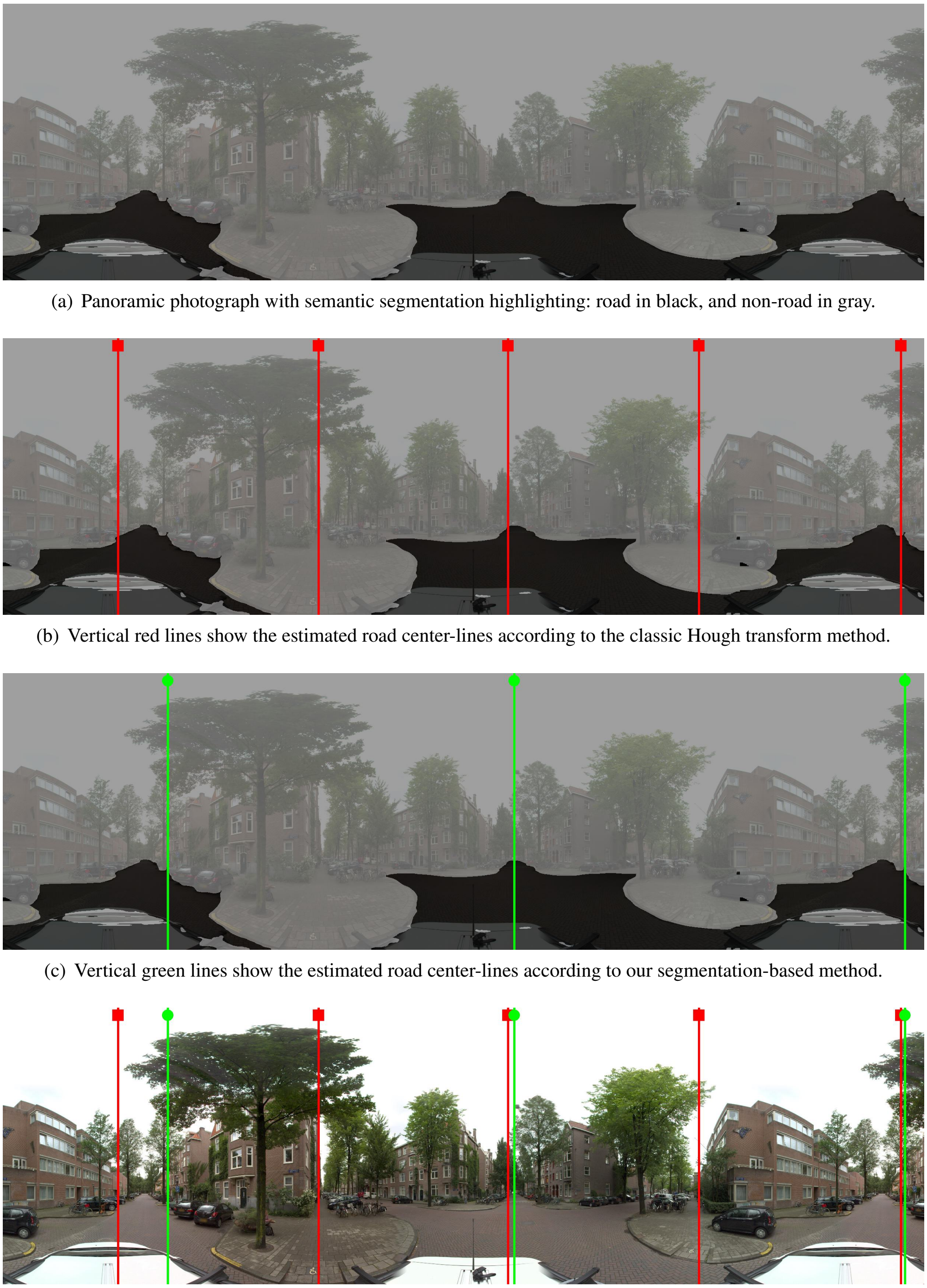
This script uses the previously obtained segmentation arrays and applies the remaining processing to crop panoramic images and determine which non-panoramic images should be discarded. The results of this script are a series of subimages from each panoramic image, saved alongside it, and a series of output logs for each image, detailing the analysis results and findings for each image. These findings can be used to select images for acceptance or rejection based on image quality. The script also outputs SQL statements for insertion into the percept-backend database (initially in a disabled state), and further statements for enabling the use of the images when ready to show to end-users.

C.4. Finding road center-lines

We show the intermediate stages of road center-line finding in Figure 7. Before that, the zeroth step (not shown) is simply to crop out the bottom quarter of the panoramic photograph because it usually only shows the vehicle holding the camera, and then to extend the panorama by wrapping the first 25% of the photo onto the right-hand side of the photo. The result will be that the first fifth of the extended photo is exactly the same as the final fifth, which can be easily seen in the given example. The reason for doing this is to prevent analysis from missing out on any boundary cases on the left or right edge of the panoramic photograph; instead we have to deal with duplicate results, which we clean up later.

We then apply semantic segmentation (Thisanke et al., 2023), in this case using the Mask2Former model (Cheng et al., 2022) trained on the CityScapes dataset (Cordts et al., 2016), readily available through the HuggingFace transformers library (Wolf et al., 2020). We are only interested in finding the portion of the image that corresponds to 'road', therefore Figure 7(a) shows the road pixels highlighted in black, and everything else is grayed out.

One of the classic road center-line finding algorithms (K et al., 2022) tries to find vanishing points in photographs, with the idea that roads tend to be linear features following perspective lines into the distance. To do this, we simply apply the venerable Hough transform (Illingworth and Kittler, 1988) as available in OpenCV (Bradski, 2000), to edges in an image and then finds the points of intersection. In Figure 7(b) we show the results of such an algorithm that has been applied to edges found (by the OpenCV Canny edge detection) in the semantic segmentation matrix, as indicated



(a) Panoramic photograph with semantic segmentation highlighting: road in black, and non-road in gray.

(b) Vertical red lines show the estimated road center-lines according to the classic Hough transform method.

(c) Vertical green lines show the estimated road center-lines according to our segmentation-based method.

(d) The original panoramic photograph overlaid with both kinds of estimated road center-line detection method results. The red lines (with squares) come from the Hough transform method, and the green lines (with circles) come from our segmentation-based method.

Figure 7: An example panorama, showing semantic segmentation and two methods of road center-line finding.

by the red lines (and tagged with squares). You can see that it finds some roads, but it also finds foot-ways. It also gets confused when near the edge of the photograph. The first and the last red lines should appear in the same relative position (to the road) because these portions of the photo are the duplicated (wrapped) first and final fifths of the photo, but in fact the lines are placed differently. This happens because some perspective lines in the first fifth, coming from the central portion of the panoramic image, do not exist in the final fifth.

Our segmentation-based method We chose a more focused method of finding center-lines. We use semantic segmentation on images to identify the pixels corresponding to roads. For each column x in the segmentation matrix, let $B(x)$ be distance between the bottom of the matrix and topmost 'road'-labeled pixel, and let $C(x)$ be the count of 'road'-labeled pixels in the bottom half of the matrix. Choosing an adjustment factor k (in our case, $k = 1/8$) then we define $R(x)$ as a combination of the above two: $R(x) = B(x) + kC(x)$. While $B(x)$ captures the intuition that roads should appear as 'peaks' in the segmentation matrix, there are sometimes spurious road pixels. Therefore, $C(x)$ ensures that there is a substantial number of road pixels behind each peak. However, $C(x)$ by itself has the problem that the camera-carrying vehicle often interferes with the segmentation results near the bottom of the image, especially around the centers of roads. Therefore we apply a scaling factor k to ensure that the sides of roads are not overemphasized in the output.

Using the Python Scipy library (Virtanen et al., 2020), we then find the peaks of $R(x)$ from left-to-right across the width of the image, under the assumption that there is one valid road to be found approximately in each third of the panorama. This relatively simple algorithm is surprisingly effective, and we found it to be more effective than the above road-finding algorithms using vanishing point perspective detection. An example is shown by the green lines (tagged with circles) in Figure 7(c).

Most notably, this method finds the centers of roads where the camera is looking directly down the road. This matches the intuitive description of what we are seeking: the view that person has when they stand in a road and look down it. Further notes: the first and third line in the example are in the same relative position, as they should be, since these sections of the image are duplicates. In both cases, the estimated center-line is a bit off from where it should be; this has been caused by spurious 'non-road' pixels from the intrusion of the camera-carrying vehicle into the image.

Figure 7(d) shows both methods overlaid onto the original panorama. In our experience, the segmentation-based method generally gave the best estimates, and even if it was off, it was not by much and the end result remained presentable. It would be better if the camera-carrying vehicle could be cut out entirely, however it does not appear in a consistent way nor is it detected consistently by semantic segmentation. Therefore, we applied a pragmatic rule of thumb and cropped the bottom quarter of the panorama, seeing that the horizon on these panoramic images always falls within the middle band to a reasonable extent.

C.5. Image quality filtering

The script `torch_process_segm.py` (see Appendix C.3) performs a number of operations for each given image, which includes computing the two key factors of our image quality algorithm: (1) 'contrast' (C) as calculated by a function derived from the Scikit-image function `skimage_contrast` (Van der Walt et al., 2014), and (2) 'tone-mapping score' (T) as described by Stefanescu (2021), which effectively tries to measure if an image is too dark, too bright or suffers from a poorly distributed range of colors as shown in a color-histogram breakdown of the image. Both C and T are numbers between 0 and 1. We define $C_{min} = 0.35$ and $T_{min} = 0.35$ as thresholds for the following tests, and $T_{floor} = 0.8$ as a threshold for an adjustment factor. Both tests must be satisfied for the image to pass our quality filter:

$$T_{min} < T$$
$$C_{min} < C + \max(0, T - T_{floor})$$

These tests effectively establish minimum contrast and tone-mapping score requirements for images. The threshold numbers were chosen by trial and error. We also adjusted the contrast test so that a particularly high tone-mapping score could compensate for a worse contrast result. We found several cases where images had slightly lower contrast than desired but had strong tone-mapping scores and they appeared reasonable to the eye. Therefore, we incorporated this adjustment factor into the contrast test so that we would not be overly pessimistic and lose those images.