

Datasheet for ELSA dataset

Maryam Hosseini^{1*} Marco Cipriano^{2*} Daniel Hodczak³
Sedigheh Eslami² Liu Liu¹ Andres Sevtsuk¹ Gerard de Melo²

¹Massachusetts Institute of Technology (MIT)

²Hasso Plattner Institute (HPI)

³University of Illinois Chicago (UIC)

maryamh@mit.edu, marco.cipriano@hpi.de

June 13, 2024

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The main motivation was the lack of annotated, still-image datasets on human activity localization to benchmark the Open-vocabulary Detection models. In response to these challenges, we propose ELSA, a new benchmark dataset and evaluation framework in order to evaluate the performance of OVD models in recognizing and localizing human activity in urban streets from still images.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was curated by Liu Liu (MIT), Maryam Hosseini (MIT), Marco Cipriano (HPI), and Sedigheh Eslami (HPI).

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and

number.

The project was funded by the HPI-MIT Designing for Sustainability program as well as the German Federal Ministry for Education and Research (BMBF) within the project KI-Servicezentrum Berlin Brandenburg (01IS22092).

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Our dataset comprises of:

1) Imagery data, from two main sources of publicly available street-view imagery: Google and Bing. Following [2], due to the limitation on disseminating of the imagery data, we only publish a metadata file in csv format with *Panoid, latitude, longitude, heading* information. We provide the code for obtaining the imagery from their respective APIs. Users need to create API key and downloading the number of images

*Equal contribution.

included in our dataset will not go beyond the free download quota on either services.

2) The segmentation masks (.png files) with sidewalk and road classes to be used for cropping the images to human-scale. Users can reconstruct the processed dataset by running our script after obtaining the imagery data. This step ensures that results are reproducible.

How many instances are there in total (of each type, if appropriate)?

ELSA includes 971 images with more than 4.3K annotated bounding boxes for social and individual activities. Out of this number, 40 images are negative samples of scenes without actual pedestrians but with printed images of people on billboards, buses, or walls for which no label is created, leaving us with a total of 931 labeled images. We provide a csv file with label data including bounding box coordinates, labels (multi-label), file ID, and 971 segmentation masks (png) per box. In total, there exists 34 distinct single labels in ELSA. Since we have a multi-labeling scheme, each bounding box can have 2 or more of the distinct 34 labels associated with it, correspondingly. As a result, ELSA includes 112 unique combinations of labels describing social interaction.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). This is a subset of all street-level images from NYC. The intent is not having an equal distribution of images of each neighborhood.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Street-level images cropped to the human scale to only contain the areas in the vicinity of the sidewalks. Following [2], due to the limitation on disseminating of the imagery data, we only publish a metadata file in csv format with *Panoid, latitude, longitude, heading* information. We provide necessary codes and reference segmentation masks for the users to create the exact dataset described here. We include the list of labels, bounding boxes, and prompts for the post-processed images in our repository.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each image is associated with a label-set (multi-label) of bounding boxes and a segmentation mask (png).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit. Yes, we will provide extensive documentation on how the images and their labels and masks are related through unique identifiers. Furthermore, there is no explicit relationship across different images.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No, the entire dataset is intended for benchmarking.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

In general, images often lack high resolution, and the capture device introduces distortion. When

it comes to detecting human activity and interactions, the task becomes even more challenging due to crowded scenes, the proximity of transportation means to sidewalks, and the anonymization process, which blurs faces and creates artifacts, complicating both labeling and detection.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

We follow prior work [2] and only release the required image metadata to be used by users to download the images from their respective API (Google street-view static and Bing Streetside). We also provide scripts to make it easier for users to obtain the data, using their own API key.

Other non-restricted street-level imagery are mostly taken by cars and looking at the road ahead, whereas, for a social interaction study we required data that looks directly at sidewalks where people congregate, and the panoramic nature of these sources provided us with the flexibility to set the camera heading to be perpendicular to the road, looking directly to the left and right side of the road, a.k.a., sidewalks.

The number of instances (images to be obtained) is lower than the free tier download quota on both Google Street View static and Bing Streetside platforms.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the

content of individuals non-public communications)? If so, please provide a description.

No. The dataset is open to public to be downloaded, however, each user must use their own API key to access and download.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

It captures street-view images that has people in it, but the faces are blurred by the distributing sources.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No, all faces all blurred by their respective distributor (Google and Bing).

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was acquired by scraping images and associated metadata from Google Street View and Bing. Each instance in the dataset corresponds to a specific location, represented by an image. Both images and metadata are provided by the Google Street View and Bing service. All data collection, curation, and filtering are done by ELSA coauthors.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data is publicly available to download via Google and Bing APIs, and this was the method to collect the data.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

We chose two neighborhoods in NYC, Manhattan known for its lively, dense, busy urban streets; and Bronx, where lower density, lower rise buildings and residential areas are prominent characteristics.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

One Ph.D. student was responsible for downloading the images, and the pre-processing was done by the team of two other PhD students and one Postdoc researcher.

Over what timeframe was the data collected? Does this timeframe match the creation time-

frame of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Google Street-View: min: 2007- 2023, median: 2014; and Bing Streetside: 2012-2018, median: 2015 - The time-frame does not have any significance in our specific use case of looking at the distribution of human interactions.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

To focus mainly on the events happening on sidewalks, we employed the CitySurfaces [1] semantic segmentation model to segment pedestrian-dedicated areas. The segmented regions were then used to crop the street-level images to only include the immediate areas around the sidewalks.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

Yes. The user will also have access to that data when first downloading the images. There are not other data that is held out and all the information of the images are provided in the meta-data shared.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes. All codes to process the data and create the cropped images will be available on our GitHub repository.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Yes. For benchmarking the Grounding DINO model described in the paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes. All codes to benchmark OVD models on our dataset is publicly available on our repository. <https://github.com/urban-submissions/elsa>

What (other) tasks could the dataset be used for?

We encourage future researchers to utilize this dataset to benchmark open-vocabulary detectors (OVDs) and action recognition/localization models, as well as for urban informatics studies to fine-tune models for social interaction detection

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No.

Are there tasks for which the dataset should not be used? If so, please provide a description.

As it stands, the dataset should be solely used for research purposes in its uncurated state. Likewise, this dataset should not be used to aid in military or surveillance tasks.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The meta data, code to download the images with the user API, pre-processing codes and masks to create the exact image dataset and the code to run all the experiments together with a documentation will be publicly available on our GitHub repository.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

GitHub will be the main platform for dissemination at this stage. Later, we will build a website for the project, with more models being added to the benchmark. The website will also enable a collaborative environment for the continued support of project. Upon official release, we will upload the relevant files to reconstruct the data to zenodo.org and include the DOI on our repository for citation.

When will the dataset be distributed?

The initial version of ELSA is available on our GitHub page. We plan to build a website around it and include a web-based documentation (read-the-doc) later in October. The dataset is already accessible on the temporary GitHub repository <https://github.com/urban-submissions/elsa>. However, ELSA will be officially distributed after the camera ready on a final repository.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

CC-BY-4.0

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

ELSA owns the metadata and release as CC-BY-4.0. We do not own the copyright of the images.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The data will be supported and maintained by both teams at MIT and HPI.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Both corresponding authors can be contacted via email maryamh@mit.edu, marco.cipriano@hpi.de. Our GitHub repository can also serve as another platform to contact the whole team.

Is there an erratum? If so, please provide a link or other access point.

There is no erratum for our initial release. Errata will be documented as future releases on the dataset GitHub repository.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The current dataset will remain unchanged unless a compelling need for an update arises. However, we anticipate future versions of the dataset, leveraging crowd-sourced street-level images that are free from licensing constraints, provided there is sufficient demand.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We will continue to support ELSA. We plan to develop a website dedicated to our project, which will serve as a platform for sharing the next versions of the data and benchmarking an expanding array of models. This website will not only showcase our current models but will also facilitate a collaborative environment, fostering ongoing support and contributions from the community. By integrating these features, we intend to create a dynamic and continually evolving resource that benefits both researchers and practitioners in the field.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Unless there are grounds for significant alteration to certain indexes, extension of the dataset will be carried out on an individual basis.

References

- [1] HOSSEINI, M., MIRANDA, F., LIN, J., AND SILVA, C. T. Citysurfaces: City-scale semantic segmentation of sidewalk materials. *Sustainable Cities and Society* 79 (2022), 103630.
- [2] NAIK, N., PHILIPOOM, J., RASKAR, R., AND HIDALGO, C. Streetscore – predicting the perceived safety of one million streetscapes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2014), pp. 793–799.