# Datasheet for StreetSurfaceVis

## I. Motivation For Datasheet Creation

### *A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)*

This dataset is intended to train machine learning models that predict surface type and quality of road parts visible in street-level images.

Typical street-level image datasets are commonly collected in good weather conditions, using only a single vehicle and camera setup within a limited geographic boundary. Perspectives of sidewalks and cycleways are usually not considered. This dataset is intended to fill the gap of heterogenous street-level image datasets containing a sufficient amount of images for each pertinent surface type and quality including roadways, bikeways and footways with varying image quality levels and perspectives, influenced by factors such as the device used, its mounting or prevailing lighting and weather conditions.

### *B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?*

Thus far, the dataset has not been used.

### *C. What (other) tasks could the dataset be used for?*

This dataset is intended to train machine learning models that predict surface type and quality of road parts visible in street-level images.

### *D. Who funded the creation dataset?*

The creation of this dataset is part of the SurfaceAI project funded by *The Federal Ministry of Transport and Digital Infrastructure (BMVI)* of Germany in the *mFUND* funding program.

### *E. Any other comment?*

The dataset was created by a research team led by Prof. Dr. Helena Mihaljević at the University of Applied Sciences, HTW Berlin.

## II. Datasheet Composition

### *A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)*

The dataset consists of street-level images from Germany gathered from the crowdsourcing platform *Mapillary*[1] with a road located in the bottom center of the image, together with a label for type and quality of the surface in focus.

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

There are 9,122 instances in total in the version V1.0.

### *C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?*

Each instance consists of an image in JPG format in four resolution levels along with metadata including the unique image id, the user id and name of contributor, the date the image was captured and the geographic location the image was taken, provided by the original resource Mapillary. The available resolution levels are the original, i.e. unprocessed, image uploaded by the contributor and three rescaled sizes with a width of 256, 1024 and 2048 pixels, respectively. The instances are associated with a label for surface type and quality.

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

Each instance has a label consisting of two values. They describe the surface type and quality of the focal road in the street-level image. The labels primarily align with the *OpenStreetMap (OSM)*[2] road segment tags `surface`[3] and `smoothness`[4], respectively. The dataset comprises images with surface type and quality that are important from a traffic perspective and represent a relevant portion of street types in Germany. This results in the type labels *asphalt*, *concrete*, *paving stones*, *sett*, and *unpaved*[5], each of which

[1] `https://www.mapillary.com/`
[2] `https://openstreetmap.org/`
[3] `https://wiki.openstreetmap.org/wiki/Key:surface`
[4] `https://wiki.openstreetmap.org/wiki/Key:smoothness`
[5] More precise options for unpaved include ground, (fine) gravel, grass, compacted, and dirt, but this level of differentiation is not relevant for our context.

accounts for at least 1% of the tagged OSM road segments in Germany. For the quality label, we restrict to five of eight proposed levels, ranging from *excellent* (suitable for rollerblades), *good* (suitable for racing bikes), *intermediate* (suitable for city bikes and wheelchairs), *bad* (suitable for normal cars with reduced velocity) to *very bad* (suitable for cars with high-clearance). Not all quality labels are suitable for all surface types.

See Table I for the number of images per each class, i.e., type-quality combination.

TABLE I
FINAL DATASET SIZE BY TYPE-QUALITY CLASS.

| | excellent | good | interm. | bad | very bad |
|---|---|---|---|---|---|
| asphalt | 971 | 1,696 | 821 | 246 | - |
| concrete | 314 | 350 | 250 | 58 | - |
| paving stones | 385 | 1,063 | 519 | 70 | - |
| sett | - | 129 | 694 | 540 | - |
| unpaved | - | - | 326 | 387 | 303 |

*E. 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 data is missing.

*F. 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.*

None.

*G. 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).*

The dataset is a sample of around 170 Million images in Germany[6] from the crowdsourcing platform *Mapillary*. The original set is not labeled with surface type and quality. The dataset is intended to represent a diverse selection of surface types and qualities, but not the original image set.

Certain images are excluded:

- images of roads with rare surface types, i.e., frequency of surface type is less than 1% for OSM road segments
- images not depicting roads (e.g., houses, rivers)
- images without a single road in the focal area (e.g., cycleway and footway are depicted equally)
- images that do not allow to tell surface type or quality, e.g., due to blurry images, dark lighting or snowy roads

By strictly limiting the number of images per geographic unit (see sampling strategy in Section III) , it is ensured that the dataset is geographically diverse.

The test data is diversified regarding spatial distribution, however, unlike training data, surface type and quality distribution are not artificially adjusted (Note, that this results in the two rare classes of *concrete - bad* and *paving stones - bad* not being present in the testset).

*H. 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.*

The dataset is split in training and testing subsets. The testset comprises 776 images from five German cities, varying in region and population size – Munich, Cologne, Lunenburg, Dresden, and Heilbronn. Test data includes geospatially distinct areas that are not present in the training data, thereby ensuring classification models' ability to generalize to unseen regions is tested.

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

Mapillary images are commonly captured in sequences of entire trips, i.e., an image is captured every few seconds. Thus, many images are commonly highly similar, as the same person with the same camera angle and weather condition takes images of the same road in short spatial distances.

To limit these redundancies the maximum number of images per sequence and location was restricted (see sampling strategy in Section III).

Due to annotation inaccuracies, incorrect labels and images with unsuitable image compositions may be included.

*J. 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.*

The dataset is entirely self-contained. All instances related to this dataset are saved, not linked. The original images and informations can be accessed on the Mapillary website[7] or via the Mapillary API[8], published by Mapillary under the

[6]Retrieved in January 2024

[7]Access of original images via `https://www.mapillary.com/app/?pKey=IMAGE_ID`

[8]`https://www.mapillary.com/developer/api-documentation`

CC-BY-SA[9] license.

## III. Collection Process

### A. 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 instances were collected via the Mapillary API, filtered (see below), and finally manual human curated and labeled. To achieve high quality labels, the authors of this dataset developed an annotation guide with example images and underwent self-organized training to manually label surface *type* and *quality*. The instructions included labeling the *focal road* located in the bottom center of the street-level image. Annotators were encouraged to consult each other for a second opinion when uncertain.

### B. 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 directly observable as raw images, except that the labels were assigned manually by human experts.

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

To ensure sufficient images per class while reducing manual annotation, three strategies to filter images were applied batch-wise. 1. Images are pre-filtered using OSM tags, i.e. geolocations of Mapillary images are spatially intersected with OSM road segments tagged with `surface` and `smoothness` and assigned with the labels of the closest OSM road segment within a maximum distance of two meters. To eliminate ambiguous street intersections, 10% of the start and end of each segment are cut off beforehand. Only images with a pre-label available in this way will be considered further. 2. Surface type classification models are iteratively trained with already curated images and applied to new batches. Only images whose OSM surface pre-label matches the type classification model prediction will be considered further. To reduce bias towards easy-to-classify examples, a random sample of 10% of the images excluded in this way is also taken into account. 3. Underrepresented type-quality combinations are amplified by prompt-based image classification with OpenAI's GPT-4o[10] model.

In addition, the number of images for each location and sequence is restricted to increase the dataset's heterogeneity. This reduces the number of images taken by the same person on one trip and thus increases spatial diversity, camera specifications, environmental conditions, and photographic perspectives. Specifically, we limit the number of images per geographic unit (*Mercantile tile* on zoom level 14[11], which is roughly equivalent to $\sim 1.5$ x $1.5km$ grid cells) to 5 and the number of images per sequence to 10 per surface type and quality class. Per class according to pre-labels, images are randomly sampled under this restriction. A target size of 300-400 images was aspired.

### D. 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)?

University employees collected and annotated the data as part of their regular working hours.

### E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe 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.

The data for the dataset was sampled from Mapillary and annotated from November 2023 to May 2024. The recording date of the images associated with the instances of the dataset ranges from 2013, the year in which Mapillary was launched, to May 2024. Note, that there are 26 outliers with an associated date before 2013.

## IV. Data Preprocessing

### A. 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.

After the application of the sampling strategies described in Section III, instances where the focus was ambiguous were manually discarded, such as when two parts of the road (e.g., the cycleway and footway) were depicted equally, or when the surface could not be classified due to factors such as snowy roads, blurry images, or non-road images.

### B. 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.

The original resource is the crowdsourcing platform Mapillary which contains all instances (see above Section II).

[9] `https://creativecommons.org/licenses/by-sa/4.0/`
[10] `https://openai.com/index/hello-gpt-4o/`
[11] We thereby adhere to the same geographic unit as utilized by the Mapillary API for computational feasibility.

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

The code used to sample and filter images is provided in this repository: `https://github.com/SurfaceAI/dataset_creation`. Note, that the code does not include the classification model used for pre-labeling of surface type.

### D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

Gerneally, yes, as initial models for surface type and quality show satisfying performance. However, some classes remain below the target size of 300 to 400 instances.

### E. Any other comments

As the focal road located in the bottom center of the street-level image is labeled, it is recommended to crop images to their lower and middle half section.

This `Python` code may be used for cropping:

```
from PIL import Image

img = Image.open(image_path)
width, height = img.size
img_cropped = img.crop((0.25*width,
                        0.5*height,
                        0.75*width,
                        height))
```

## V. Dataset Distribution

### A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The dataset is published here:
`https://doi.org/10.5281/zenodo.11449977`

### B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset was first released in June 2024. It is distributed under the CC-BY-SA license.

### C. Are there any copyrights on the data?

According to the Mapillary License[12], the copyright of each image remains with its contributor.

Generally, the terms of use of Mapillary apply[13].

### D. Are there any fees or access/export restrictions?

None.

## VI. Dataset Maintenance

### A. Who is supporting/hosting/maintaining the dataset?

The research team of the project SurfaceAI[14].

### B. Will the dataset be updated? If so, how often and by whom?

Generally, the dataset does not require updates to remain useful for its intended purpose. However, there may be future updates enriching data with additional attributes, adding further surface classes, enhancing underrepresented classes, or refining annotations.

### C. How will updates be communicated? (e.g., mailing list, GitHub)

Updates will be displayed on the dataset repository and the project website.

### D. If the dataset becomes obsolete how will this be communicated?

In the unlikely event of becoming obsolete, this will be displayed on the dataset repository and the project website.

### E. Is there a repository to link to any/all papers/systems that use this dataset?

None.

(Citations of the dataset DOI may provide respective information, e.g., using Google Scholar)

### F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Contributions to the dataset are welcome. Please contact `surface-ai@htw-berlin.de` for respective inquiries.

## VII. Legal and Ethical Considerations

### A. 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.

Not applicable.

[12] `https://help.mapillary.com/hc/en-us/articles/115001770409-Licenses`

[13] `https://www.mapillary.com/terms`

[14] `https://surfaceai.github.io/surfaceai/`

*B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

No.

*C. 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.

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

The data set does not primarily relate to people, but some of the images depict people or license plates. The non-identifiability is described below. The individual contributor user names of the original dataset are indicated for each image.

*E. 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.

*F. 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.*

People who are depicted in the images are made unrecognizable by the original provider (Mapillary) through blurring[15] and are only available in this way. As every image in this dataset is human-curated, images would have been excluded and reported if otherwise. The contributors of the images to the original dataset are indicated by their user name and ID for each image in this dataset, as required by the license of the original dataset. As this data is already provided in the original dataset, which is openly accessible, this dataset does not publish any new data that could be used to identify individuals.

*G. 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.*

As the images in the dataset also contain vehicles, license plates are shown which are made unrecognizable by the original provider through blurring and are only available in this way. As every image in this dataset is human-curated, images would have been excluded and reported if otherwise.

[15]See Section 6 in `https://www.mapillary.com/terms`

*H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*

As described above, the data was collected from the crowdsourcing platform Mapillary.

*I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

No, the contributors were not notified. However, they actively uploaded images for the purpose of data sharing and thereby agreed to the Mapillary terms of use and thus, the sharing of images under the CC-BY-SA license.

*J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

Yes, by uploading images they agreed to the Mapillary terms of use and thus, the sharing of images under the CC-BY-SA license.

*K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

Not applicable.

*L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

Not applicable.