Data in Brief 69 (2026) 113407

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

TaPaD: An instance-level image dataset for tactile paving with deterioration annotations

Yuto Matsuda, Yuki Matsuda*

Okayama University, 3-1-1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama, 700-8530, Japan

ARTICLE INFO

Article history:
Received 21 August 2026
Revised 18 October 2026
Accepted 31 October 2026
Available online 3 November 2026

Keywords:
Object detection
Visual impairment
Accessibility
Road infrastructure
Smart city

ABSTRACT

Tactile paving provides navigational cues for people with visual impairments, but its surfaces can deteriorate through cracking, soiling, fading, heat deformation, and wear. This data article presents TaPaD, an instance-level road-surface image dataset for tactile-paving detection, block-type classification, and deterioration assessment. TaPaD contains 1611 images captured with a bicycle-mounted compact camera and 3192 images captured with smartphones. All bicycle images are annotated, comprising 6506 tactile-paving instances. Of the smartphone images, 1200 are annotated, comprising 13,962 instances; the remaining 1992 images are released as unannotated images. Each annotated instance includes a bounding box, one block-type label (guiding or warning), and zero or more of five deterioration labels (cracked, dirty, faded, melting, and worn-out). Per-image metadata include image dimensions, timestamps, and GPS coordinates where available. The repository contains an annotation guideline and Python scripts for conversion to YOLO and multi-hot formats, while the dedicated TaPaT annotation tool is publicly available separately. Annotation agreement was assessed using 20 images independently annotated by four additional evaluators, with bounding-box count agreement, mean intersection over union, Jaccard index, and κ statistics reported. The dataset can be reused for object detection, multi-label

* Corresponding author.
E-mail address: yukimat@okayama-u.ac.jp (Y. Matsuda).
Social media: @cocolabJP (Y. Matsuda)

https://doi.org/10.1016/j.dib.2026.113407
2352-3409/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

classification, infrastructure inspection, and exploratory analysis of images collected under the two acquisition settings.
© 2026 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Specifications Table

Subject	Computer Sciences
Specific subject area	Object detection, type classification, and deterioration assessment of tactile paving
Type of data	Road-surface images depicting tactile paving (JPG and PNG); instance-level annotations (JSON); metadata (CSV); annotation guideline (PDF); agreement-assessment outputs (CSV, PDF, and TXT); conversion and evaluation scripts (Python)
Data collection	Road-surface images were captured using a bicycle-mounted compact camera (1611 images) and iPhone 13 and iPhone 15 Pro smartphones through the Repot web application (3192 images). All bicycle images and 1200 smartphone images were annotated at the instance level with bounding boxes, one block-type label, and zero or more deterioration labels. Agreement was assessed on 20 images annotated by four additional evaluators.
Data source location	City/Town/Region: Primarily Okayama Prefecture, with additional images from 17 other prefectures in Japan (Fukuoka, Oita, Ehime, Kochi, Kagawa, Hiroshima, Hyogo, Osaka, Nara, Kyoto, Shizuoka, Toyama, Tokyo, Kanagawa, Hokkaido, Okinawa, and Tottori) and eight other countries (South Korea, Germany, Italy, Spain, Belgium, the United Kingdom, the Netherlands, and Sweden). Country: Japan (primary), with minor subsets from the eight countries listed above. GPS coordinates: Per-image coordinates are provided where available.
Data accessibility	Repository name: Zenodo Data identification number: 10.5281/zenodo.23388177 Direct URL to data: https://doi.org/10.5281/zenodo.23388177 Instructions for accessing these data: Download the published files (approximately 3.7 GB compressed) and see "README.md" for details of the directory structure and usage.
Related research article	Y. Matsuda and Y. Matsuda, "Tactile Paving Detection and Classification Method Based on Cyclist-Participatory Road Image Sensing," Companion Proceedings of the 27th International Conference on Distributed Computing and Networking, pp. 78–83, 2026. https://doi.org/10.1145/3737611.3776614.

1. Value of the Data

- TaPaD contains 20,468 tactile-paving instances annotated with one block-type label and zero or more deterioration labels, supporting object-detection, block-type classification, and multi-label deterioration tasks.
- The instance-level deterioration annotations extend the reuse of tactile-paving imagery beyond pedestrian guidance and navigation to applications such as automated infrastructure inspection and maintenance-data analysis.
- The dataset combines images collected by a bicycle-mounted compact camera in Okayama Prefecture and by handheld smartphones across 18 prefectures in Japan, providing examples captured from different viewpoints and under different acquisition settings. A limited number of smartphone images from eight other countries are included as supplementary references to road environments outside Japan.
- All 3192 smartphone images are released, including 1992 unannotated images that can be used for self-supervised, semi-supervised, active learning, or additional annotation studies.
- The released annotation guideline, separately available TaPaT annotation tool, and conversion scripts support extension of the dataset and reuse in YOLO-format object detection and multi-hot deterioration classification workflows.

2. Background

According to the World Health Organization, at least 2.2 billion people worldwide have a vision impairment or blindness [1], and approximately 273,000 people in Japan were living with visual impairment in fiscal year 2022 [2]. Tactile paving supports safe and independent mobility and has been standardized in Japan and internationally [3,4]. However, surveys in Japan have documented installation and maintenance deficiencies, while on-foot inspection limits the area that can be monitored routinely [5,6]. Existing tactile-paving datasets mainly support detection, tracking, or segmentation [7–9] and do not provide deterioration labels for individual paving units. The related research article [10] described the bicycle-based sensing system and a tactile-paving detection and classification method but did not publicly release the underlying image dataset. TaPaD provides the collected images, instance-level block-type and deterioration annotations, metadata, annotation guideline, and conversion tools for public reuse.

3. Data Description

The TaPaD dataset comprises road-surface images and corresponding instance-level annotations of tactile paving and is hosted on Zenodo [11]. The repository structure is summarized in Table 1.

3.1. Image files

This dataset consists of road-surface images captured using two acquisition methods. The bicycle subset contains 1611 JPG images at 1280×1024 pixels, of which 178 include GPS coordinates. The smartphone subset contains 3192 PNG images at 1920×1080 pixels, all with GPS coordinates. All images in both subsets are publicly released.

3.2. JSON annotation files

The “annotations/” directory contains one JSON file for each annotated image. All 1611 bicycle images are annotated, corresponding to 1611 JSON files. Of the 3192 smartphone images, 1200 are annotated, corresponding to 1200 JSON files. The remaining 1992 smartphone images are released without annotation files.

Each JSON file contains image-level metadata and a list of instance-level annotations. The schema is summarized in Table 2. Furthermore, each bounding box corresponds to one tactile-paving unit and is annotated according to the criteria defined in “annotation_guideline.pdf”.

3.3. CSV metadata files

Each “metadata.csv” serves as an index covering all images in the corresponding subset (bicycle or smartphone), including both annotated and unannotated images. Each row corresponds to one image and provides the following fields: “filename”; “width” and “height” (image dimensions in pixels); “timestamp” (ISO 8601 format, UTC); and “latitude” and “longitude” (GPS latitude and longitude at the time of capture). For images without available GPS information, “latitude” and “longitude” are left blank.

“metadata.csv” is intended to support image search based on capture time or location, independent of annotation status. For images that have been annotated, the corresponding annotation file can be retrieved by replacing the file extension of “filename” with “.json” (see “annotations/”).

Table 1
Dataset directory structure and organization.

Directory/File	Description	Contents
bicycle/images	Road-surface images captured with a bicycle-mounted compact camera (1280×1024)	1611 JPG images (71 MB)
bicycle/annotations	Annotations corresponding to "bicycle/images"	1611 JSON files (one per image)
bicycle/metadata.csv	Per-image metadata for the bicycle image subset	6 columns: filename, width, height, timestamp, latitude, longitude
smartphone/images	Road-surface images captured with handheld smartphone cameras (1920×1080)	3192 PNG images (3.3 GB)
smartphone/annotations	Annotations corresponding to "smartphone/images"	1200 JSON files (one per annotated image); the remaining 1992 images are released without annotation files
smartphone/metadata.csv	Per-image metadata for the smartphone image subset	6 columns: filename, width, height, timestamp, latitude, longitude
tools/tapad2yolo.py	Converts TaPaD JSON annotations to YOLO-format labels	Python script
tools/tapad2multihot.py	Converts TaPaD JSON annotations to multi-hot-encoded labels	Python script
evaluation/selected_images	Assessment images and sampling log	20 images (10 bicycle and 10 smartphone) and sampled_log.txt
evaluation/annotations/{M,A,B,C,D}	Initial annotations used for the agreement assessment	Five subdirectories, each containing 20 JSON files
evaluation/matching_results/M_vs_{A-D}	Pairwise matching results between the primary annotator M and each additional evaluator	Four subdirectories; each contains bbox_comparison.csv, bbox_comparison.pdf, and statistics_summary.txt
evaluation/matching_results/A_vs_B_vs_C_vs_D	Multi-evaluator matching results among evaluators A–D	bbox_comparison.csv, bbox_comparison.pdf, and statistics_summary.txt
evaluation/sample_images.py	Selects assessment images using the weighted sampling procedure	Python script
evaluation/match_annotations.py	Matches bounding boxes and exports comparison results	Python script
evaluation/calculate_statistics.py	Calculates the agreement statistics from the matching results	Python script
README.md	Dataset documentation	Dataset description
LICENSE	License terms governing reuse of the dataset	CC BY 4.0 license text
annotation_guideline.pdf	Full annotation criteria for bounding-box delineation, block-type classification, and deterioration labeling	PDF document

3.4. Python tool files

The "tools/" directory contains two Python scripts that convert TaPaD annotations into formats directly usable by common machine learning frameworks.

"tapad2yolo.py" converts TaPaD JSON annotations into YOLO-format [12] label files (class ID together with normalized center coordinates, width, and height). Class IDs are assigned to

Table 2
JSON annotation schema.

Field	Description
image.filename	Image file name
image.width, image.height	Image dimensions in pixels
image.gps.lat, image.gps.lon	GPS coordinates at capture time; null if unavailable
image.timestamp	Capture timestamp
annotations[].bbox	Bounding box in [x, y, width, height] format (top-left pixel coordinates)
annotations[].attributes.block_type	Block type: "guiding" or "warning"
annotations[].attributes.deteriorations	List of zero or more deterioration labels: "cracked", "dirty", "faded", "melting", "worn-out"

"guiding" and "warning", allowing the converted labels to be used directly with standard object detection frameworks.

"tapad2multihot.py" converts the deterioration state of each tactile-paving instance into a multi-hot-encoded CSV file with columns "filename", "cracked", "dirty", "faded", "melting", and "worn-out". Each row corresponds to one block instance, with 1 indicating the presence and 0 the absence of each deterioration label. With the corresponding option, the script also saves a cropped image for each bounding box. When multiple instances originate from one source image, a two-digit sequential number is appended to the source filename stem (e.g., "<source>_00.jpg", "<source>_01.jpg", ...). The generated crop filename is also written to the "filename" column of the CSV. The output can be used to prepare inputs for multi-label classification frameworks such as PyTorch.

3.5. Evaluation files

The "evaluation/" directory contains the materials used for the annotation agreement assessment. The "selected_images/" subdirectory contains the 20 assessment images (10 bicycle and 10 smartphone images) together with "sampled_log.txt", which records the sampling settings and the label distribution of the selected images. The "annotations/" subdirectory contains the initial annotations from the primary annotator (M) and four additional evaluators (A–D), organized in separate subdirectories.

The "matching_results/" subdirectory contains the outputs of bounding-box matching for the four M-versus-evaluator comparisons ("M_vs_A" to "M_vs_D") and the comparison among the four additional evaluators ("A_vs_B_vs_C_vs_D"). Each comparison directory contains a CSV table of matched and unmatched annotations, a PDF visualization of the matching results, and "statistics_summary.txt". The CSV files also contain the deterioration labels recorded after guideline clarification and reassessment in fields named "after_survey" or "<annotator>_after_survey". The scripts "sample_images.py", "match_annotations.py", and "calculate_statistics.py" implement the assessment-image sampling, annotation matching, and agreement-statistics calculations reported in this article, respectively.

3.6. Label definitions and overall statistics

Each bounding box has one block-type label and zero or more deterioration labels. Table 3 reports both the number of images containing at least one instance of each label and the number of annotated instances. Because one image can contain multiple instances and deterioration annotations are multi-label, the image-count and deterioration-category columns are not mutually exclusive. The statistics for the smartphone subset in Table 3 are based only on the 1200 man-

Table 3
Number of images containing each label and number of annotated instances by subset.

Label	Bicycle		Smartphone		Total	
	Images	Instances	Images	Instances	Images	Instances
guiding	1522	5246	1152	8370	2674	13,616
warning	236	1260	478	5592	714	6852
cracked	81	118	740	2341	821	2459
dirty	1107	3953	971	8988	2078	12,941
faded	89	338	111	815	200	1153
melting	15	32	48	219	63	251
worn-out	56	104	314	1405	370	1509
No deterioration	560	2153	508	2690	1068	4843

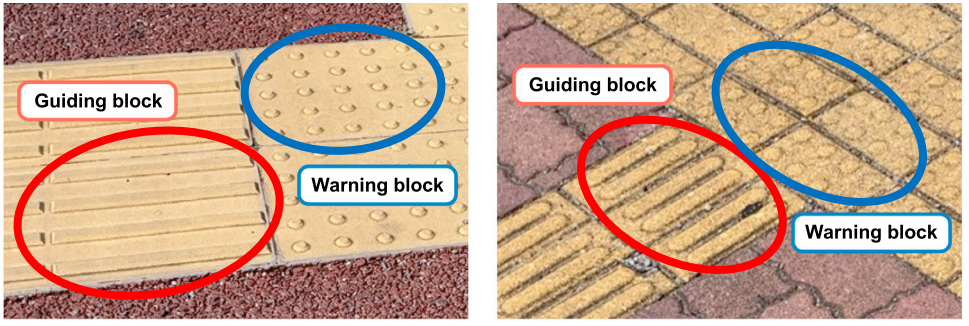


Fig. 1. Examples of tactile paving block types (guiding and warning).

ually selected annotated images, and the observed label frequencies should not be extrapolated to the full set of 3192 smartphone images.

The block-type label is either “guiding” or “warning”. Guiding blocks indicate the direction of travel, while warning blocks consist of an array of protrusions intended to alert pedestrians. Examples are shown in Fig. 1.

The deterioration annotation is multi-label and consists of zero or more of five categories: “cracked”, “dirty”, “faded”, “melting”, and “worn-out”. The absence of all five labels indicates no visually detectable deterioration. Because multiple deterioration labels can be assigned to one instance, the deterioration-label counts in Table 3 are not mutually exclusive. Definitions are provided below, and examples are shown in Fig. 2.

- cracked: cracking, partial loss of material, or discontinuities caused by manholes or pavement joints, limited to cases where continuous installation would otherwise have been possible. Loss limited to the protrusions alone is classified as “worn-out”, whereas loss of protrusions accompanied by surrounding material loss is classified as “cracked”.
- dirty: overall blackening, localized staining, or dark speckled soiling of the block surface.
- faded: overall whitening of the block, insufficient contrast with the surrounding pavement, or a marked reduction in saturation relative to neighboring blocks.
- melting: visually apparent deformation or flattening of the protrusions consistent with heat-related softening.
- worn-out: abrasion of the protrusions, partial loss of protrusions without surrounding material loss, or exposure of the underlying ground due to abrasion of the pavement.

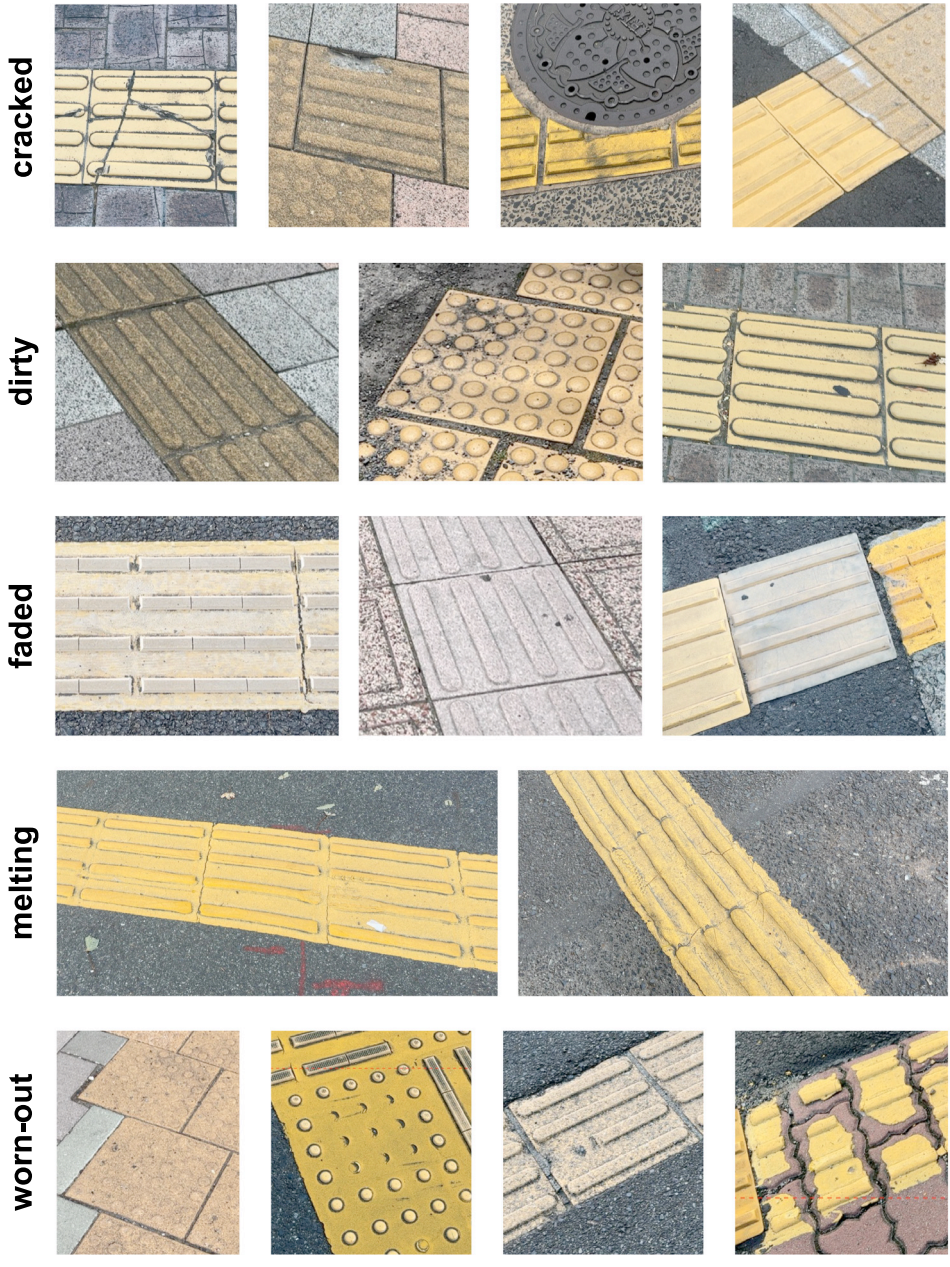


Fig. 2. Examples of deteriorated tactile paving.

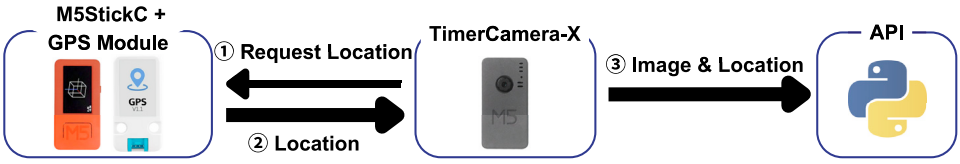


Fig. 3. System architecture of BraiLoop.

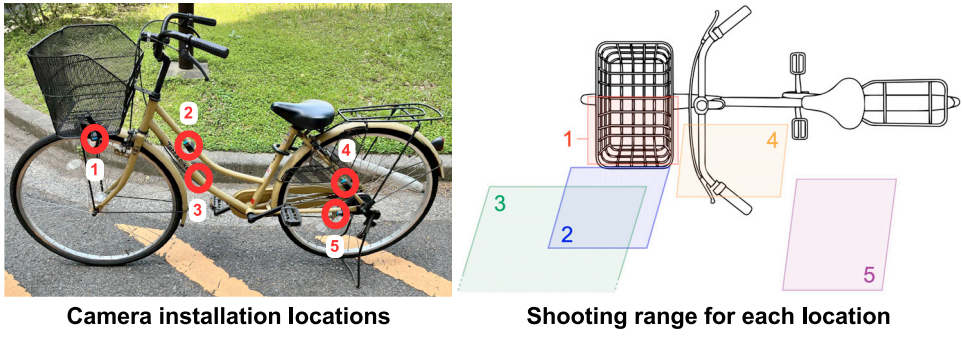


Fig. 4. Camera installation locations and their shooting ranges.

4. Experimental Design, Materials and Methods

4.1. Data collection

4.1.1. Bicycle images

Bicycle images were collected using "BraiLoop", a participatory sensing system consisting of a compact camera (M5Stack TimerCamera-X) and a GPS module (M5Stack GPS Unit v1.1) mounted on a bicycle [10]. Fig. 3 shows the system architecture. Images were captured automatically while riding, with the next capture initiated immediately after an image was captured and transmitted to the API. The prototype supports a capture rate of up to about 10 images per minute.

The camera was mounted at five positions shown in Fig. 4. Over a period of approximately two months, a total of 11,014 road-surface images were collected within Okayama City. From these images, those containing visible tactile paving were selected, and images in which motion blur prevented reliable identification of block boundaries or type were excluded, yielding a final set of 1611 images for annotation.

A GPS module was integrated into the system partway through data collection. Consequently, 1433 images captured before this integration do not include GPS information, while the remaining 178 images, captured afterward, do. Fig. 5 shows a map of the collected images; for images without GPS information, the traveled route is indicated with a red line.

4.1.2. Smartphone images

Smartphone images were collected using "Repot" [13], a participatory geo-tagged photo collection platform implemented as a web application that automatically attaches location information to each submitted photograph. Photographs were captured with an iPhone 13 or iPhone 15 Pro. Images tagged in Repot as containing tactile paving were extracted, and all 3192 extracted images were included in the dataset without additional quality-based exclusion. All images in this subset include GPS coordinates. The authors collected the images over approximately 15 months.

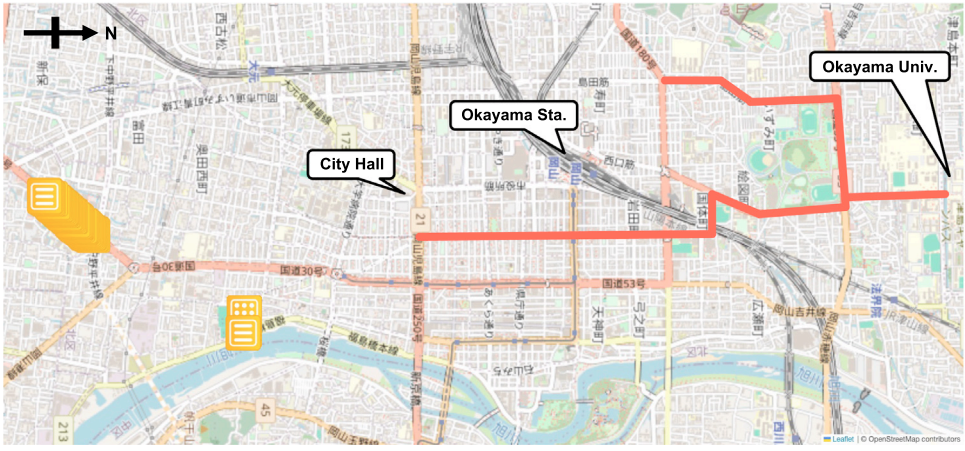


Fig. 5. Geographic distribution of bicycle images (in Okayama City).

Table 4
Distribution of smartphone images by prefecture in Japan and by country.

Japan Prefecture	Number of images	Country	Number of images
Okayama	1728	Japan	3144
Kochi	401	South Korea	17
Nara	205	Spain	8
Osaka	193	Italy	7
Tokyo	170	Germany	6
Okinawa	120	Sweden	5
Hokkaido	68	United Kingdom	3
Toyama	66	Belgium	1
Hiroshima	52	Netherlands	1
Fukuoka	41		
Tottori	27		
Hyogo	18		
Ehime	17		
Kanagawa	15		
Oita	8		
Kagawa	8		
Kyoto	6		
Shizuoka	1		
Total	3144	Total	3192

Of the 3192 smartphone images, 1200 were manually selected for annotation to broadly cover the geographic regions represented in the full smartphone subset. No formal random or stratified sampling procedure was applied. Accordingly, the annotated subset should not be interpreted as a random or statistically representative sample of all 3192 smartphone images, and the observed label frequencies in the 1200 annotated images should not be extrapolated to the full smartphone collection.

Images were captured in 18 prefectures in Japan, predominantly in Okayama Prefecture, and eight other countries. Nevertheless, the geographic distribution is strongly concentrated in Japan (3144/3192 images), particularly in Okayama Prefecture (1728 images); the images from the eight other countries are small supplementary subsets rather than balanced international coverage. The distributions by prefecture and country are shown in Table 4. Fig. 6 maps the image locations.

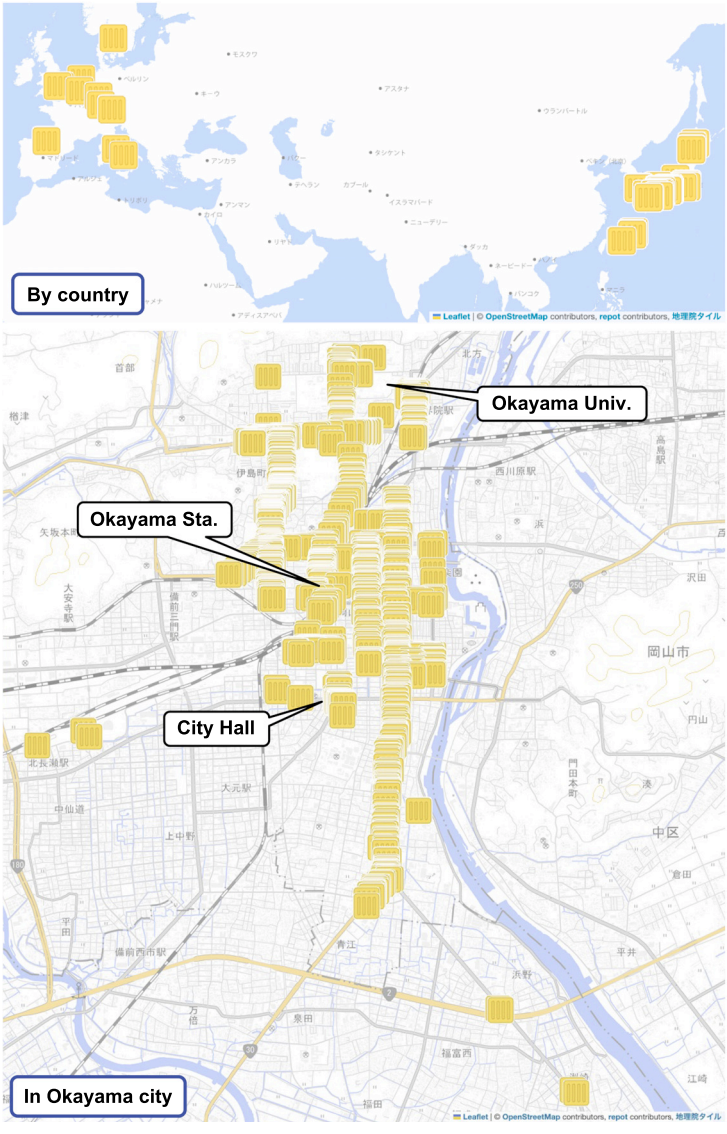


Fig. 6. Geographic distribution of smartphone images (by country and in Okayama City).

Note that these two subsets were collected under different acquisition and selection pipelines and are intended to provide acquisition diversity, not a controlled comparison between bicycle-mounted and smartphone imaging.

4.2. Annotation procedure

Bounding-box, block-type, and deterioration annotations were produced by a single primary annotator (Yuto Matsuda; M). Annotation was initially performed using “Label Studio” [14], a general-purpose open-source annotation tool. However, to enable the assignment of multiple

labels (deterioration labels) to a single bounding box and to improve annotation efficiency for this task, a purpose-built annotation tool, "TaPaT" [15], was later developed and adopted for the remainder of the annotation work.

Most annotations were produced manually. After an initial subset of images had been annotated, a YOLO11-based detection model [10] trained on that subset was used to generate candidate bounding boxes for subsequent images. The primary annotator manually reviewed and corrected all candidates and assigned the block-type and deterioration labels. The released metadata do not record whether each image was annotated entirely manually or initialized from model-generated proposals. The same annotation guideline was applied to manually created and model-assisted annotations.

4.3. Annotation quality assessment

4.3.1. Assessment procedure

To describe variation associated with single-annotator labeling, an annotation agreement assessment was conducted on a subset of the dataset. The assessment involved the primary annotator (M) and four additional evaluators without prior annotation experience (A–D) and used 20 images (10 bicycle and 10 smartphone images). The four additional evaluators read the annotation guideline before performing the initial annotation task.

After the initial annotation task, a guideline-clarification and reassessment session was conducted. For all matched bounding boxes, the primary annotator and the additional evaluators assigned identical block-type labels. For each differing deterioration label, M explained the applicable guideline criterion and the visual evidence in the image, after which the evaluator reassessed the label. Evaluators could also delete a bounding box when it was judged not to correspond to an annotatable tactile-paving unit. No bounding boxes were added, and the coordinates of retained bounding boxes were not moved or resized. The initial results therefore quantify agreement from independent annotation, whereas the reassessed results quantify agreement after guideline-based clarification.

4.3.2. Selection of assessment images. Images used in the assessment were selected from images previously annotated by the primary annotator using a weighted random sampling procedure based on a set-cover heuristic. At the time of the agreement assessment, 1000 smartphone images had been annotated and constituted the candidate pool for assessment-image selection. The annotated smartphone subset was subsequently expanded to 1200 images for the released dataset. At each step t, let $\mathcal{R}^{(t)}$ denote the set of remaining candidate images, L_i denote the set of labels present in image i (block type and deterioration categories), $U^{(t)}$ denote the set of labels not yet covered by images selected before step (t), $c_l^{(t)}$ denote the number of previously selected images containing label l, and D_i denote the number of distinct deterioration categories present in image i. The score assigned to each candidate image $i \in \mathcal{R}^{(t)}$ is

$$s_i^{(t)} = 2\left|L_i \cap U^{(t)}\right| + 0.3\max\left(D_i - 1,\ 0\right) - 0.1\sum_{l \in L_i} c_l^{(t)} \tag{1}$$

The three terms represent, respectively, coverage of previously uncovered labels, a bonus for images containing multiple deterioration types, and a penalty discouraging redundant selection of labels that were already well represented. Using this score, the selection probability for candidate image i is

$$P\left(i|\mathcal{R}^{(t)}\right) = \frac{\max\left(s_i^{(t)},\ 0.01\right)^{\tau}}{\sum_{j \in \mathcal{R}^{(t)}} \max\left(s_j^{(t)},\ 0.01\right)^{\tau}} \tag{2}$$

One image is drawn without replacement according to this distribution; $U^{(t)}$and $c_l^{(t)}$ are then updated, and the procedure is repeated until 10 images have been selected. The parameter τ



Fig. 7. Selected images for the annotation quality assessment: 10 images from the bicycle subset (top two rows) and 10 images from the smartphone subset (bottom two rows).

Table 5
Number of annotated instances by label in the images selected for the annotation quality assessment.

Label	Bicycle	Smartphone	Total
guiding	30	97	127
warning	21	15	36
cracked	3	38	41
dirty	32	42	74
faded	11	6	17
melting	2	23	25
worn-out	1	35	36
No deterioration	6	26	32

controls the sharpness of the selection-probability distribution: smaller values yield a more uniform distribution, whereas larger values concentrate probability mass on higher-scoring candidates and approach greedy selection. A value of $\tau = 1.5$ was used.

The procedure was applied independently to the bicycle and smartphone subsets, yielding 10 images from each subset (20 images in total). The selected images are shown in Fig. 7, and their label distribution is summarized in Table 5. This label-coverage-oriented sample was used to examine all categories and was not intended to reproduce the natural label distribution of the full dataset.

4.3.3. Evaluation method

The evaluation used two comparisons. The first compared the primary annotator M with each of the four additional evaluators (M–A, M–B, M–C, and M–D). The second compared the four additional evaluators with one another to describe variation among evaluators without prior annotation experience.

Metrics included per-image bounding-box (BBox) count agreement, mean IoU (mIoU) for matched bounding boxes, the Jaccard index for deterioration-label sets, Cohen's κ per deterioration label for the M-versus-evaluator comparisons, and Fleiss' κ per deterioration label for the four-evaluator comparison. For each image, bounding-box count agreement was defined as $\min(n_1, n_2)/\max(n_1, n_2)$, where n_1 and n_2 are the numbers of bounding boxes in the two annotations, and was then averaged across images. mIoU and the Jaccard index were calculated for

matched box pairs, averaged within each image, and then averaged across images. When both deterioration-label sets were empty, the Jaccard index was defined as 1. κ statistics were calculated by pooling matched bounding boxes across images and were then averaged over the five deterioration labels. Unmatched boxes were excluded from the mIoU, Jaccard, and κ calculations, although their presence affected bounding-box count agreement. During reassessment, bounding boxes could be deleted but were not added, moved, or resized; therefore, mIoU for retained matched boxes was unchanged and is reported once.

4.3.4. Matching annotations across multiple annotators

Bounding boxes from M and each additional evaluator were matched one-to-one using the Hungarian algorithm with negative IoU as the cost. To determine the IoU threshold for candidate matches, multiple threshold values were tested on images for which the annotators produced different numbers of bounding boxes, and the resulting assignments were compared with correspondences determined by visual inspection. A threshold of 0.3 was selected because it produced no discrepancies from the visually determined correspondences in these cases.

For matching bounding boxes among the four evaluators, one-to-one matching via the Hungarian algorithm was performed for each of the six pairs of evaluators, and the resulting pairwise matches were transitively merged using a union-find (disjoint-set) structure to group detections presumed to correspond to the same tactile paving unit. This procedure enables all evaluators to be treated symmetrically, without designating any one as a reference. When transitive merging resulted in more than one detection from the same annotator being assigned to a single group, only the detection with the highest summed IoU to the other members of the group was retained, and the remaining detections were separated into independent groups. This algorithm is presented in Algorithm 1.

After applying the matching procedure with an IoU threshold of 0.3, all resulting matches and unmatched bounding boxes for every assessment image were visually inspected. We confirmed that the algorithmic matching results were consistent with the manually determined correspondences in all assessment images. Detections corresponding to the same tactile-paving unit were confirmed as paired, and boxes present in only one evaluator's annotations were retained as unmatched. An example is shown in Fig. 8.

4.3.5. Descriptive evaluation results

Agreement metrics between M and each evaluator are shown in Tables 6 and 7. Here, in Tables 6–9, Pre denotes the initial independent annotation, and Post denotes reassessment after guideline clarification. However, the Post values should not be interpreted as independent inter-annotator reproducibility; they quantify agreement after guideline-based clarification by the primary annotator. Because the 20 assessment images were selected to cover all label categories rather than to reproduce their natural prevalence, the agreement statistics reported in Tables 6–9 should be interpreted as descriptive results for this evaluation set rather than as population-level estimates for the entire TaPaD dataset. Initial bounding-box count agreement ranged from 0.84 to 0.91, and mIoU ranged from 0.72 to 0.81. The mean initial Cohen's κ across deterioration labels ranged from 0.28 to 0.47. At the label level, initial κ for cracked ranged from 0.66 to 0.87, whereas initial κ for melting ranged from -0.07 to 0.41.

After guideline clarification and reassessment, mean Cohen's κ across deterioration labels ranged from 0.80 to 0.94, and the mean deterioration-label Jaccard index across the M-versus-evaluator comparisons was 0.91. For evaluator C, κ for melting was 0.13 after reassessment, while the corresponding values for evaluators A, B, and D were 1.00, 0.98, and 0.97, respectively.

Agreement metrics among the four additional evaluators are shown in Table 8, with label-level Fleiss' κ and pairwise Cohen's κ in Table 9. Before clarification, bounding-box count agreement was 0.86, mIoU was 0.75, the deterioration-label Jaccard index was 0.50, and mean Fleiss' κ across the five labels was 0.35. The mean Cohen's κ for the four M-versus-evaluator comparisons was 0.39.

Algorithm 1 Multi-annotator bounding-box grouping via transitive matching.

```
Input: D = [D[0], ..., D[N-1]], where D[a] is the detection list of annotator a
       IoU threshold theta
Output: G, a list of groups represented as {annotator_index: local_detection_index}

Function MatchPairs(A, B, theta)
  if A is empty or B is empty then return []
  for i <- 0 to len(A)-1 do
    for j <- 0 to |B|−1 do I[i,j] <- IoU(A[i].bbox, B[j].bbox)
  (R, C) <- HungarianAssignment(-I)
  M <- []
  for k <- 0 to |R|−1 do
    if I[R[k], C[k]] ≥ theta then M.append((R[k], C[k], I[R[k], C[k]]))
  return M

Function Find(parent, x)
  if parent[x] != x then parent[x] <- Find(parent, parent[x])
  return parent[x]

Procedure Union(parent, x, y)
  rx <- Find(parent, x); ry <- Find(parent, y)
  if rx != ry then parent[rx] <- ry

Function ClusterAnnotations(D, theta)
  N <- len(D); gid <- {}; nextID <- 0
  for a <- 0 to N-1 do
    for i <- 0 to |D[a]|−1 do gid[a,i] <- nextID; nextID <- nextID + 1
  parent[g] <- g for g <- 0 to nextID-1; W <- {}

  for each annotator pair (a,b), a < b do
    for each (i,j,s) in MatchPairs(D[a], D[b], theta) do
      p <- gid[a,i]; q <- gid[b,j]
      Union(parent, p, q); W[p,q] <- s; W[q,p] <- s

  C <- empty map from root ID to member list
  for each (a,i) in Keys(gid) do
    p <- gid[a,i]; C[Find(parent,p)].append((a,i,p))

  G <- []
  for each member list c in Values(C) do
    A <- empty map from annotator index to entries
    for each (a,i,p) in c do A[a].append((i,p))
    core <- {}
    for each a in Keys(A) do
      if len(A[a]) = 1 then
        core[a] <- A[a][0].local_index
      else
        best <- argmax_(i,p) in A[a] sum W[p,q] over (b,j,q) in c with b != a,
          where an undefined W[p,q] is treated as 0
        core[a] <- best.local_index
        for each (i,p) in A[a] except best do G.append({a:i})
    G.append(core)
  return G
```

After clarification and reassessment, mean Fleiss' κ across the five labels was 0.87. Label-level Fleiss' κ was 0.98 for cracked, 0.90 for dirty, 0.95 for faded, 0.65 for melting, and 0.89 for worn-out. For melting, the three evaluator pairs not involving C had post-clarification Cohen's κ values of 0.97–1.00, whereas the three pairs involving C had values of approximately 0.14–0.15.

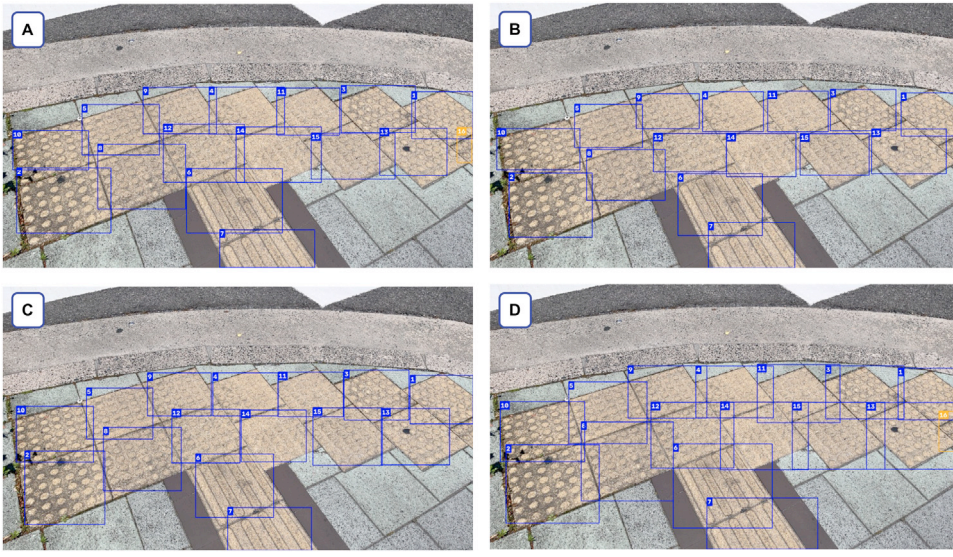


Fig. 8. Example of annotation matching results.

Table 6
Agreement between the primary annotator (M) and each additional evaluator (A–D) at Pre and Post. Values in brackets indicate the range across images.

	BBox count agreement		mIoU	Deterioration-label Jaccard index		Annotation time (min)
	Pre	Post		Pre	Post	
A	0.84 [0.20, 1.00]	1.00	0.76 [0.37, 0.95]	0.57 [0.00, 1.00]	0.89 [0.07, 1.00]	66.2
B	0.91 [0.50, 1.00]	0.98 [0.75, 1.00]	0.81 [0.62, 0.91]	0.43 [0.00, 1.00]	0.91 [0.50, 1.00]	41.4
C	0.88 [0.50, 1.00]	0.99 [0.89, 1.00]	0.81 [0.63, 0.94]	0.58 [0.00, 1.00]	0.93 [0.36, 1.00]	65.7
D	0.86 [0.50, 1.00]	1.00	0.72 [0.56, 0.84]	0.27 [0.00, 0.78]	0.91 [0.57, 1.00]	47.5
Mean	0.87	0.99	0.78	0.47	0.91	55.2

*Annotation time for the primary annotator (M) was 17.5 min.

4.3.6. Descriptive summary

Before guideline clarification, the mean Cohen's κ across the four M-versus-evaluator comparisons was 0.39, and the mean Fleiss' κ among the four additional evaluators was 0.35. These values reflect agreement from independent annotation. After clarification and reassessment, the corresponding mean values increased to 0.90 and 0.87, indicating that many disagreements could be resolved through guideline-based clarification rather than reflecting independent reproducibility. Melting had the lowest post-clarification Fleiss' κ among the five labels (0.65).

The primary annotator completed the 20-image task in 17.5 min. Completion times for evaluators A–D were 66.2, 41.4, 65.7, and 47.5 min, respectively, with a mean of 55.2 min. These times describe the assessment task and are provided as metadata on annotation effort.

Table 7
Cohen's κ for deterioration labels between the primary annotator (M) and each additional evaluator (A–D) at Pre and Post.

	cracked		dirty		faded		melting		worn-out		Mean	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
A	0.84	0.97	0.41	0.80	0.36	1.00	0.00	1.00	0.54	0.95	0.43	0.94
B	0.66	0.97	0.53	0.84	0.21	1.00	0.41	0.98	−0.02	0.85	0.36	0.93
C	0.87	1.00	0.54	0.93	0.51	1.00	0.05	0.13	0.39	0.92	0.47	0.80
D	0.72	0.97	0.25	0.91	0.17	0.87	−0.07	0.97	0.33	0.87	0.28	0.92
Mean	0.77	0.98	0.43	0.87	0.31	0.97	0.10	0.77	0.31	0.90	0.39	0.90

Table 8
Agreement among the four additional evaluators (A–D) at Pre and Post.

Metric	Pre	Post
BBox count agreement	0.86 [0.30, 1.00]	0.98 [0.75, 1.00]
mIoU	0.75 [0.60, 0.85]	N/A[a]
Deterioration-label Jaccard index	0.50 [0.17, 0.75]	0.94 [0.71, 1.00]

[a] Post-assessment mIoU was not recalculated because retained bounding boxes were not moved or resized.

Table 9
Pairwise Cohen's κ and Fleiss' κ among the four additional evaluators (A–D) at Pre and Post.

		cracked		dirty		faded		melting		worn-out		Mean	
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Pairwise Cohen's κ	A, B	0.78	0.98	0.24	0.81	0.15	1.00	0.00	0.97	0.16	0.84	0.26	0.92
	A, C	0.71	1.00	0.31	0.98	0.27	1.00	0.00	0.15	0.21	0.93	0.30	0.81
	A, D	0.85	0.98	0.20	0.97	0.11	0.90	0.00	0.97	0.34	0.87	0.30	0.94
	B, C	0.62	0.98	0.57	0.83	0.53	1.00	−0.02	0.14	0.35	0.91	0.41	0.77
	B, D	0.77	0.96	0.48	0.84	0.53	0.90	0.48	1.00	0.33	0.89	0.52	0.92
	C, D	0.64	0.98	0.32	0.98	0.49	0.90	−0.02	0.14	0.28	0.87	0.34	0.78
Fleiss' κ	A, B, C, D	0.73	0.98	0.32	0.90	0.32	0.95	0.12	0.65	0.28	0.89	0.35	0.87

Limitations

The bounding-box, block-type, and deterioration annotations were produced by a single primary annotator; therefore, systematic annotator bias cannot be excluded. The agreement assessment used 20 images selected to cover the label categories rather than to reproduce the natural label distribution of the full dataset. Because the 1200 annotated smartphone images were manually selected with geographic coverage in mind, the annotated subset should not be interpreted as a random sample of all 3192 smartphone images. The post-clarification measurements describe reassessment after guideline-based explanation and are not an independent replication of the initial annotations. The deterioration labels represent visually inferred surface conditions and should not be interpreted as engineering-grade diagnoses of material deterioration or its physical causes. The melting label is relatively infrequent (251 instances) and showed the lowest and most variable agreement among the deterioration categories. These results indicate that particular care is required when applying the melting criterion and should be taken into account when interpreting model performance for this category. GPS coordinates are available for only 178 of the 1611 bicycle images because the GPS module was introduced during data collection. All 3192 smartphone images are released, but only 1200 currently have instance-level annotations; the remaining 1992 are provided as unannotated images. Most images were collected in Japan, particularly Okayama Prefecture, and the images from other countries form small subsets.

Ethics Statement

The 4803 publicly released images depict road surfaces captured in public spaces. Before release, all images, including the 1992 unannotated smartphone images, were visually inspected for recognizable faces and vehicle license plates, and no such personally identifiable information was identified.

The annotation agreement assessment involved four additional evaluators, who were informed of the task in advance and provided consent to participate.

The dataset construction did not involve animal experiments or data collected from social media platforms. The authors confirm that they have read and comply with the ethical requirements for publication in Data in Brief.

CRediT Author Statement

Yuto Matsuda: Conceptualization, Methodology, Software, Data Curation, Investigation, Writing – Original Draft, Visualization. **Yuki Matsuda:** Conceptualization, Supervision, Writing – Review & Editing, Visualization.

Acknowledgements

The authors acknowledge four additional evaluators for their time and effort in annotating the evaluation samples used to assess annotation agreement.

This study was supported in part by JSPS KAKENHI Grant Number JP24K20763 and JST COI-NEXT Grant Number JPMJPF2115.

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.

References

[1] World Health Organization, WHO launches first World report on vision, 2019 (accessed 17 July 2026). https://www.who.int/news-room/detail/08-10-2019-who-launches-first-world-report-on-vision .

[2] Ministry of Health, Labour and Welfare, 2022 Survey on difficulties in daily life, etc. (Nationwide Survey of Persons with Disabilities Living at Home). https://www.mhlw.go.jp/toukei/list/seikatsu_chousa_r04.html, 2022 (accessed 17 July 2026, in Japanese).

[3] Japanese Standards Association, JIS T 9251:2014 guidelines for older persons and persons with disabilities — shapes, dimensions and patterns of raised parts of tactile walking surface indicators for persons who are blind or with seeing impairment. https://webdesk.jsa.or.jp/preview/pre_jis_t_09251_000_000_2014_e_ed10_i4.pdf, 2014 (accessed 17 July 2026).

[4] International Organization for Standardization, ISO 23599: 2019 Assistive products for blind and vision-impaired persons — tactile walking surface indicators, 2019 (accessed 17 July 2026). https://www.iso.org/standard/76106.html .

[5] Ministry of Internal Affairs and Communications, Survey on the maintenance and management of tactile paving for the visually impaired, 2018 (accessed 17 July 2026, in Japanese). https://www.soumu.go.jp/main_content/000547116.pdf .

[6] Ministry of Internal Affairs and Communications, Appropriate installation and maintenance of tactile paving and related facilities for the visually impaired, 2023 (accessed 17 July 2026, in Japanese). https://www.soumu.go.jp/main_content/000901004.pdf .

[7] T. Takano, T. Nakane, J. Yu, C. Zhang, Tactile paving detection and tracking using Tenji10K dataset, IEEJ Transact. Electric. Electr. Eng. 19 (10) (2024) 1661–1672, doi:10.1002/tee.24123.

[8] X. Zhang, L. Liang, S. Zhao, Z. Wang, GRFB-UNet: a new multi-scale attention network with group receptive field block for tactile paving segmentation, Expert. Syst. Appl. 238 (2024) 122109, doi:10.1016/j.eswa.2023.122109.

[9] W. Li, Y. Song, Z. Jin, Y. Wu, Z. Zhang, Y. Xie, C. Wang, G. He, TactPav: a vision-language annotated multi-modal dataset for tactile paving navigation, Pattern Recognit. Comp. Vis. (2026) 384–398, doi:10.1007/978-981-95-5761-5_27.
[10] Y. Matsuda, Y. Matsuda, Tactile paving detection and classification method based on cyclist-participatory road image sensing, in: Companion Proceedings of the 27th International Conference on Distributed Computing and Networking, 2026, pp. 78–83, doi:10.1145/3737611.3776614.
[11] Y. Matsuda, Y. Matsuda, TaPaD: an instance-level image dataset of deteriorated tactile paving (v2.5), Zenodo (2026), doi:10.5281/zenodo.23388177.
[12] G. Jocher, J. Qiu, Ultralytics YOLO11 (v11.0.0) [Computer software], 2024. https://github.com/ultralytics/ultralytics.
[13] Y. Matsuda, Repot: a participatory geo-tagged photo collection platform, in: The 15th International Conference on Mobile Computing and Ubiquitous Networking (ICMU '25), 2025, pp. 1–2, doi:10.23919/ICMU65253.2025.11219151.
[14] M. Tkachenko, M. Malyuk, A. Holmanyuk, N. Liubimov, Label Studio: data labeling software [Computer software], 2025. https://github.com/HumanSignal/label-studio.
[15] Y. Matsuda, TaPaT: tactile paving annotation tool (v1.0.0) [Computer software], 2026. https://github.com/cocolabJP/TaPaT.