A global dataset of continuous urban dashcam driving

Md Shadab Alam[1,*], **Olena Bazilinska**[2], **Pavlo Bazilinskyy**[1]

[1]Eindhoven University of Technology, Eindhoven, The Netherlands
[2]National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

[*]Corresponding author: Md Shadab Alam (m.s.alam@tue.nl)

Abstract

We introduce CROWD (City Road Observations With Dashcams), a manually curated dataset of ordinary, minute scale, temporally contiguous, unedited, front facing urban dashcam segments screened and segmented from publicly available YouTube videos. CROWD is designed to support cross-domain robustness and interaction analysis by prioritising routine driving and explicitly excluding crashes, crash aftermath, and other edited or incident-focused content. The release contains 51,753 segment records spanning 20,275.56 hours (42,032 videos), covering 7,103 named inhabited places in 238 countries and territories across all six inhabited continents (Africa, Asia, Europe, North America, South America and Oceania), with segment level manual labels for time of day (day or night) and vehicle type. To lower the barrier for benchmarking, we provide per-segment CSV files of machine-generated detections for all 80 MS-COCO classes produced with YOLOv11x, together with segment-local multi-object tracks (BoT-SORT); e.g. person, bicycle, motorcycle, car, bus, truck, traffic light, stop sign, etc. CROWD is distributed as video identifiers with segment boundaries and derived annotations, enabling reproducible research without redistributing the underlying videos.

Keywords: Dashcam video dataset; Naturalistic driving; Urban driving; Object detection; Multi-object tracking; YouTube

1 Background & Summary

Urban traffic is a visually and behaviourally complex environment in which vulnerable road users (VRUs), particularly pedestrians and cyclists, interact with motor vehicles under substantial variation in infrastructure, lighting, road geometry, and culturally specific driving norms. The World Health Organisation estimates around 1.19 million deaths from road traffic per year, with pedestrians accounting for about 21% and cyclists approximately 5% of fatalities worldwide [1]. Automated transport research therefore benefits from datasets that capture VRU appearance and motion across diverse urban settings, rather than within a narrow set of locations or carefully curated scenarios.

Police reported collision statistics, such as the UK's STATS19 system [2], and national crash databases, such as the US Fatality Analysis Reporting System (FARS) [3], provide essential and standardised measures of crash outcomes and circumstances. STATS19 captures around 50 coded data elements per injury collision (e.g. time and location, vehicle types and manoeuvres, and basic driver and casualty attributes), while FARS codes more than 170 data elements per fatal crash spanning crash level, vehicle and driver level, and person level variables (e.g. injury severity, restraint use, and alcohol involvement). However, these resources are typically released as *tabular* records, meaning that each crash (and its associated vehicles and people) is represented as rows of coded variables rather than as continuous, time aligned video or other sensor streams. They are also centred on the crash event and its immediate circumstances. Consequently, they rarely preserve the continuous visual context needed to study how interactions unfold over time, how occlusion and scene clutter affect detection, or how routine exposure relates to risk. Naturalistic driving studies address part of this limitation by instrumenting volunteer vehicles to record extended video from the driver's viewpoint together with vehicle state over long periods, but they are expensive to conduct and often constrained by privacy and access restrictions, which limits reuse and benchmarking at scale [4–7].

Curated benchmarks for driving perception, spanning general street scene driving datasets such as KITTI [8], Cityscapes [9], ApolloScape [10], BDD100K [11], Mapillary Vistas [12], KITTI 360 [13], and A2D2 [14], as well as autonomous vehicle focussed multi sensor datasets such as Argoverse [15], Argoverse 2 [16], nuScenes [17], the Waymo Open Dataset [18], and PandaSet [19], have accelerated methodological advances by providing high quality sensor data and annotations. These benchmarks are fundamental for detection, segmentation, tracking, and forecasting, but they also reflect practical constraints of collection. Geographic coverage is often limited to a small number of cities or regions, many releases prioritise short selected scenarios rather than long uninterrupted driving, and fleets instrumented with specialised sensor suites do not necessarily reflect the viewpoint, mounting geometry, or image characteristics of widely deployed camera based driver assistance systems. VRU-focused datasets further support pedestrian and cyclist modelling, but are also geographically bounded. EuroCity Persons is explicitly European, recorded in 31 cities across 12 European countries [20]. PIE was recorded in the centre of Toronto, ON, Canada [21]. JAAD was recorded in a limited set of locations in North America and Europe [22]. Models trained on one dataset can generalise poorly to new data distributions, reinforcing the need for diverse and ecologically valid data to reduce dataset bias and domain shift [23].

Publicly available web video offers an alternative route to scale and diversity, especially via dashcam recordings. However, web shared dashcam content is frequently shaped by selection effects. Clips that are posted and widely circulated disproportionately feature unusual or high conflict events, including collisions, near misses, and confrontations, which skews the observed distribution away from routine driving. This bias is visible in dashcam benchmarks that focus on accidents and critical events [24]. For research questions that require representative samples of everyday urban driving, including typical VRU exposure, everyday interactions, and background traffic flow, there remains a gap for geographically diverse data deliberately filtered towards ordinary continuous driving in urban settings.

A dataset captured from a forward facing dashcam provides advantages for both method development and behavioural analysis[1]. It matches the egocentric viewpoint used by many deployed camera based pipelines and it foregrounds perceptual challenges that dominate real deployments, including partial occlusion by parked vehicles and street furniture, dense roadside clutter, frequent scale changes as VRUs approach the ego vehicle, and complex intersection geometries that produce ambiguous motion cues. Dashcams are widespread in many regions and are typically mounted in broadly comparable windscreen positions, enabling scalable collection across regions while maintaining a coherent observational geometry for cross locality and cross country comparisons[2]. However, prevalence and motivations are country specific, with dashcams used mainly for safety and evidential purposes in some contexts and more often for leisure recording and online sharing in others.

The CROWD (City Road Observations With Dashcams) dataset addresses these needs by curating extended duration, front facing urban dashcam video segments sourced from publicly available online recordings, with an emphasis on ordinary continuous driving. Videos containing crashes, crash aftermath, or other incident-focused content are excluded by design. The dataset prioritises routine urban conditions across multiple countries and localities and, to our knowledge, provides the largest geographic coverage among publicly available curated datasets of ordinary, temporally contiguous urban dashcam driving. It also offers a large total retained duration, particularly within this broad geographic scope. CROWD supports studies of robustness and domain generalisation, as well as analyses that depend on temporal continuity, such as multi object tracking, exposure estimation, and interaction characterisation. Alongside video identifiers, segment definitions, and dataset metadata, CROWD provides machine generated object bounding boxes aligned to video frames as a derived data product, enabling reproducible baselines and facilitating follow on experiments in detection, tracking, and interaction analysis without requiring users to rerun the full detection pipeline.

[1] Dashcam cameras may be mounted inside the vehicle (for example behind the windscreen) or externally (for example on the roof or bonnet). Some externally mounted setups can produce footage that looks very similar to an in vehicle dashcam because of the camera angle and settings.

[2] Throughout this paper, we use the terms *country* and countries to refer to countries and territories as defined by ISO 3166, including dependent territories, represented using ISO 3166 1 alpha 3 codes (field `iso3`) [25].

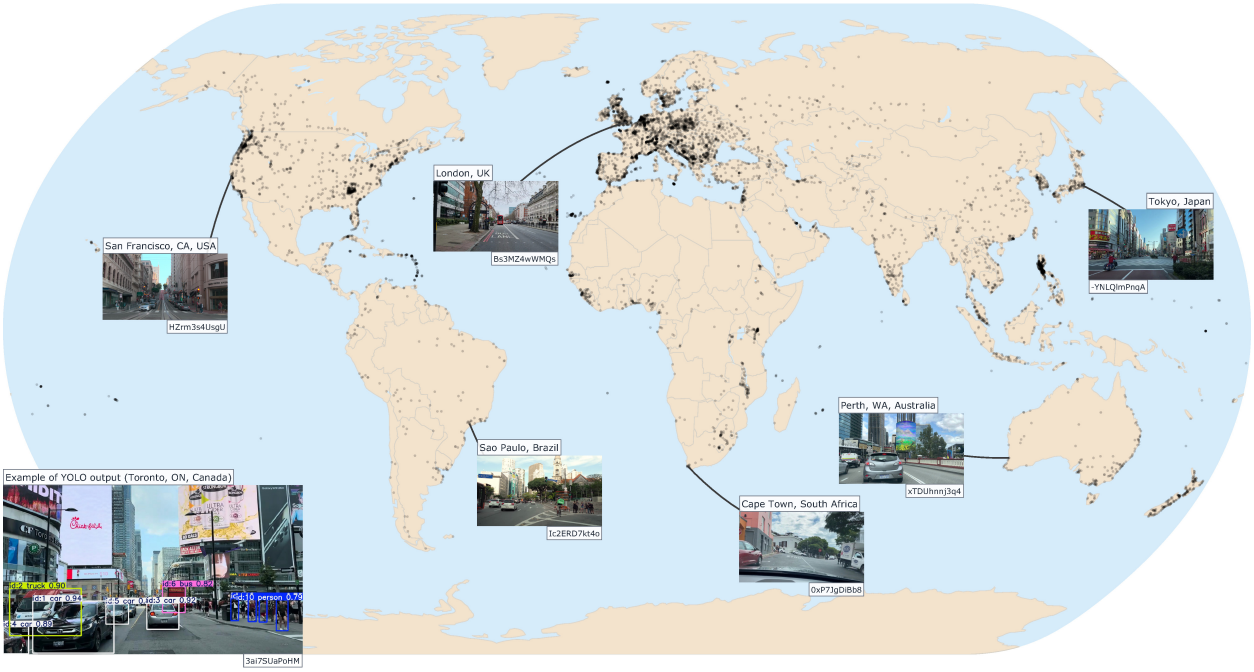

Figure 1: Geographic coverage of CROWD across 7,103 localities worldwide. Each marker denotes a locality with at least one retained dash cam segment. Marker opacity reflects the total retained footage in hours, with darker markers indicating more footage. Insets show example localities linked to their map locations.

1.1 Related datasets

A substantial body of prior work has released driving video datasets to support perception, tracking, and scene understanding. Although early driving video resources date back to CamVid [26] and the Caltech Pedestrian Dataset [27], large scale dataset releases have accelerated over the last decade as in vehicle cameras have become inexpensive, storage has become cheaper, and online video platforms have made recording and sharing at scale more accessible. This shift has been reinforced by the widespread use of dashcams for evidential purposes in collisions and disputes, including insurance related claims and fraud attempts, and by sharing clips to seek help from others or to support reporting via trusted authorities such as the police, although concerns about privacy, retaliation, and criminal misuse can shape what is uploaded and retained [28, 29]. These datasets vary in objective, acquisition pipeline, sensor suite, geographic coverage, and the extent to which they capture urban routine driving as continuous video sequences rather than isolated frames. Table 1 summarises representative resources that provide a video stream mounted on the vehicle in the forward direction, offering a viewpoint that is broadly comparable to consumer dashcam footage even when the sensor package or mounting differs.

Several of the resources summarised in Table 1 are distributed as clips or log segments on the order of tens of seconds, for example 20 s scenes in nuScenes [17] and 15 to 30 s sequences in Argoverse [15]. Such durations are well suited to benchmarking per frame detection, segmentation, and related tasks, but offer limited continuity for analyses that require longer continuous observation, such as interaction and yielding behaviour over extended encounters, exposure estimation expressed per hour of driving, and cross national behavioural measures derived from event timing in naturalistic footage [30]. BDD100K, for instance, provides 100,000 front facing 40 s clips with diversity in weather and time of day, yet the fixed clip length constrains longer horizon interaction analysis and exposure oriented measures [11]. D^2-City similarly provides urban dashcam clips with substantial scenario diversity, but the data released are segmented into short excerpts [31]. Comparable design choices appear in short log releases from multisensor platforms, where the forward camera stream is available but the temporal window per scene is brief, as in nuScenes, Waymo Open Dataset, Argoverse 2 Sensor, and PandaSet [16–19]. These datasets are invaluable for model development

and evaluation, while their temporal granularity can limit studies that require uninterrupted driving context.

Table 1: Representative driving datasets reported in the literature that provide a forward facing camera stream. For multi sensor datasets, the comparison refers to the forward camera stream or streams only. The cited papers are publicly accessible. Dataset access may be subject to registration, approval, or a licence agreement, and availability can change over time. Footage duration is reported in hours when explicitly stated by the source. Otherwise, the closest published proxy, such as distance or frame counts, is listed and indicated via footnote.[3]

Dataset	**Geographic coverage (countries/localities)**	**Footage (h)**	**Day/Night**	**Notes (dashcam video only)**
100 Car Naturalistic Driving Study (Phase II) [4]	USA (Northern Virginia, Metro Washington, DC)	≈47,383[a]	Both	Instrumented vehicles with five camera views including a forward view; infrared cabin lighting supports night driving capture; event database includes crashes, near crashes, and incidents; a public release is reported for an event subset (time series and annotations).
4Seasons [32]	Germany (locations not itemised)	N/A[b]	Both	Vehicle mounted forward view available; reported primarily by distance rather than total hours.
A2D2 [14]	Germany (Gaimersheim, Ingolstadt, Munich)	N/A[c]	Not specified	Multi sensor dataset; sequential camera frames are available (reported as 392,556 frames across three cities).
A3CarScene [33]	Italy (Marche region)	31	Not specified	Two dashcams mounted on front and rear windows; recorded on public roads (audio also provided).
Argoverse 2 (Sensor) [16]	USA (6 cities: Austin, TX; Detroit, MI; Miami, FL; Palo Alto, CA; Pittsburgh, PA; Washington, DC)	≈4.2 to 5.6[d]	Both	1,000 short logs (15 to 20 s) with surround cameras; forward facing stereo pair available; multi sensor logs.
BDD100K [11]	USA (New York, NY; San Francisco Bay Area, CA; other regions not specified)	≈1111[e]	Both	100,000 front view videos; each is ∼40 s; includes diverse weather and times of day (daytime and nighttime).
Boreas [34]	Canada (Toronto, ON)	N/A[f]	Both	Multi sensor dataset; forward camera stream available; reported primarily by distance (>350 km).
CADC [35]	Canada (Region of Waterloo, ON)	N/A[g]	Not specified	Adverse winter driving dataset; forward camera streams available (8 cameras); reported primarily by frame counts.
Caltech Pedestrians [27]	USA (Los Angeles, CA)	10	Not specified	Forward facing video recorded from a moving vehicle in urban traffic; annotated pedestrian bounding boxes with occlusion labels; available via CaltechDATA.
comma2k19 [36]	USA (CA–280 between San Jose and San Francisco, CA)	>33	Not specified	Road facing camera; 2,019 segments of 1 minute each.
D^2–City [31]	China (6 cities, not itemised in the source)	∼100	Not specified	Front facing dashcam clips; the released collection is about one hundred hours.
DR(eye)VE [37]	Not stated (urban, countryside, motorway routes)	≈6.2[h]	Both	Roof mounted car perspective video; 74 sequences of 5 minutes each; recorded at daytime and at night (also includes weather variation).
HDD (Honda) [38]	USA (San Francisco Bay Area, CA)	104	Not specified	Instrumented vehicle dataset; this row refers only to the front facing stream (dashcam style viewpoint).
IDD–CRS[39]	India (unstructured traffic; 30 day collection)	90	Not specified	5,400 untrimmed front view videos (1 min each) collected via dashcam; long tail critical road scenarios.
IDD–X [40]	India (Hyderabad region; urban, rural, motorway)	85	Both	Dual view (front and rear) driving videos; captured across day and night and varied weather; front view is dashcam style.

Continued on next page.

Table 1 (continued).

Dataset	Geographic coverage (countries/localities)	Footage (h)	Day/Night	Notes (dashcam video only)
KITTI (raw) [8]	Germany (Karlsruhe)	6	Day	Vehicle mounted stereo camera; front facing viewpoint; raw driving sequences.
KITTI 360 [13]	Germany (Karlsruhe)	N/A[i]	Not specified	Multi sensor platform; forward facing frames available; published primarily via distance and frame counts (for example 73.7 km) rather than total hours.
nuPlan [41]	USA (Boston, MA; Pittsburgh, PA; Las Vegas, NV), Singapore	1,200[j]	Not specified	Human driving dataset with multi sensor logs; forward camera stream available.
nuScenes [17]	USA (Boston), Singapore	≈5.6[k]	Both	1,000 scenes × 20 s; 6 cameras (front included) + LiDAR/radar; includes night and rain.
ONCE [42]	Not fully enumerated (200 km^2 driving regions)	144	Both	1M LiDAR scenes with 7M camera images; 7 cameras + LiDAR; diverse environments (day, night, sunny, rainy, urban, suburban).
OpenDV 2K (OpenDriveLab) [43]	≥40 countries, ≥244 cities	2,059 (1,747 YouTube)	Not quantified	Web mined front view driving videos paired with text; country and city counts are estimates from video titles; camera setup is described as uncalibrated.
Oxford RobotCar [44]	UK (Oxford)	N/A[l]	Both	Repeated route over ∼1,000 km; data collected across varied weather and lighting conditions; multi camera platform, using the forward view stream(s) as dashcam style video.
PandaSet [19]	USA (California: Silicon Valley, San Francisco)	≈0.23[m]	Both	Multi sensor dataset with forward camera stream; 103 scenes of 8 s each.
PhysicalAI-AV (NVIDIA) [45]	25 countries, 2,500+ cities	1,727	Both	306,152 clips × 20 s; 7 RGB cameras including front wide and tele; access gated by NVIDIA AV dataset licence.
Waymo Open Dataset (Perception) [18]	USA (San Francisco, CA; Phoenix, AZ; Mountain View, CA)	≈11.3[n]	Both	2030 segments × 20 s; multi camera + LiDAR (forward view available); diverse conditions including night and varied weather.
ZOD (Zenseact Open Dataset) [46]	14 European countries	9.7	Both	Overall dataset spans 14 countries; dashcam video includes 1,473 sequences of 20 s (8.2 h) and 29 drives totalling 1.5 h.

[a]Hours of driving data collected reported as 47,382.65 h in the Phase II report, rounded to ≈47,383 h. [b]4Seasons is reported primarily by distance and sequences rather than total hours. [c]A2D2 is reported via sequential frame counts across cities. [d]1000×(15 to 20) s ≈ 4.2 to 5.6 h. [e]100,000×40 s ≈ 1,111 h. [f]Boreas is reported primarily by distance (350+ km) rather than total hours. [g]CADC is reported primarily by frame counts (for example 7000 frames annotated) rather than total hours. [h]74×5 min ≈ 6.17 h. [i]KITTI 360 is published primarily via distance and frame counts, rather than total hours (for example 73.7 km). [j]nuPlan reports 1,200 h of human driving data. [k]1000×20 s ≈ 5.56 h. [l]Oxford RobotCar is reported as distance and images rather than total hours. [m]103×8 s ≈ 0.23 h. [n]2030×20 s ≈ 11.28 h. [o]1473×20 s ≈ 8.18 h (Sequences only).

Other related datasets emphasise repeated routes, long term variation, or challenging conditions, typically within a small number of operating areas. Oxford RobotCar provides repeated traversals of a fixed route under various conditions, including night, but the data are tied to a single city and are often summarised by distance or image counts rather than a simple hour based measure [44]. Related long term resources such as 4Seasons [32] and Boreas [34] target seasonal and weather variation on repeated routes, again prioritising controlled revisit structure over broad cross city coverage. Datasets such as CADC [35] focus on adverse conditions and winter driving, providing forward camera streams alongside other sensors, with a geographical scope restricted to a limited region.

[3]Access status checked on 4 March 2026. We could verify a public access route for 23 of the 26 datasets listed, via open download, registration, or acceptance of a licence agreement. IDD-CRS is currently listed as private on India Data, HDD requires a university affiliated download request, and ZOD requires requesting access by email.

A further set of benchmarks relies on instrumented vehicles and multi sensor acquisition to provide carefully captured urban driving sequences with extensive annotations. KITTI, KITTI-360, and A2D2 have been crucial to progress in autonomous driving research, but their geographic scope is limited to a small number of cities and collection campaigns [8, 13, 14]. ONCE provides large scale multi sensor driving data reported by regions rather than a global set of cities [42]. HDD and DR(eye)VE incorporate forward video alongside additional signals and support driver behaviour and attention related analyses, rather than functioning as globally distributed dashcam corpora [37, 38]. Comma2k19 provides a road facing a stream on a fixed commute route, which is useful for longitudinal studies on a single corridor, but does not represent a wide variation in urban life [36]. Recent resources such as IDD-CRS and IDD-X provide important coverage of unstructured traffic in India, but remain geographically concentrated relative to a cross country design goal [39, 40].

Web sourced collections can expand geographic breadth at scale by leveraging publicly hosted driving videos. OpenDV 2K reports broad coverage across at least 40 countries and at least 244 cities, illustrating the potential of web hosted material for diversity, while also inheriting heterogeneity in camera configuration and variability in metadata quality [43]. Recent work on driving world models has highlighted the scalability of single view methods based on monocular ego video and the use of weakly curated Internet sourced driving footage for model adaptation [47]. PhysicalAI-AV provides large scale multi camera data in many locations, with access governed by a dataset licence [45]. Web scale datasets can therefore provide breadth, yet they are not necessarily curated to prioritise routine urban driving segments over atypical events or non driving intervals.

Taken together, these design choices leave limited support for studies that require broad geographic coverage, routine urban driving, and longer continuous temporal context within a single resource. CROWD is intended to complement existing datasets by providing manually filtered, minute scale contiguous clips of ordinary urban driving from publicly available videos. In addition to the raw video segments, CROWD provides frame aligned object bounding boxes generated by a documented pipeline, together with structured data records and technical validation to support reproducible baselines.

2 Methods

2.1 Video source and selection criteria

A large volume of dashcam driving footage is available online, with YouTube (`https://www.youtube.com`) being a dominant hosting platform. Network traffic measurements report that YouTube accounts for about 16% of downstream volume on fixed networks and about 21% on mobile networks. [48]. Much widely shared dashcam content is selected for entertainment value and often concentrates on atypical or high conflict events (for example, crashes, near misses, and confrontations), exemplified by curated sharing channels such as the Telegram group BadShofer (`https://t.me/s/badshofer`), 40,716 subscribers, accessed 31 March 2026). Such clips are typically short, incident focused, and frequently include changes in viewpoint or editing, making them poorly suited for analyses of routine urban exposure.

In contrast, a separate genre on YouTube consists of long, continuous recordings captured with relatively stable, forward facing equipment (often longer than 40 minutes per upload; for example: `https://www.youtube.com/@jutah`, approximately 834,000 subscribers and 894 videos, accessed 31 March 2026). These videos are commonly consumed for their relaxing, monotonous visual and auditory characteristics and can trigger Autonomous Sensory Meridian Response (ASMR) [49–51] or serve as ambient background viewing [52]. Previous work on ASMR on YouTube identified driving themed content as a recurring theme within ASMR videos [53]. We targeted the latter category to curate footage representative of routine urban driving.

2.2 Search strategy, screening, segmentation, and manual labels

Candidate videos were identified manually by the authors through YouTube search using phrases such as *dashcam driving in [city]*, *driving video in [city]*, *dashcam videos in cities*, and *dashcam driving in [country]*. Throughout this paper, we use *locality* to denote the named inhabited place associated with a record, including hamlets, villages, towns and cities [54]. All identification, screening, and segmentation decisions

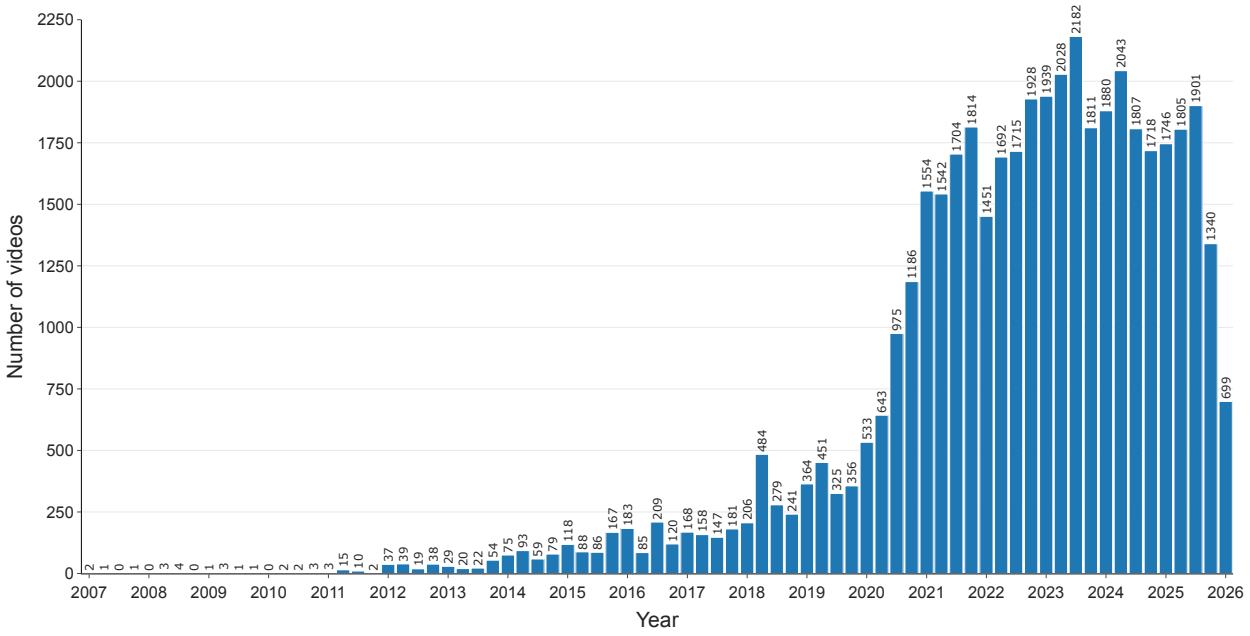

Figure 2: Upload date distribution of YouTube videos contributing at least one retained CROWD segment. Bars show the number of videos per calendar quarter, based on YouTube upload date metadata. Quarters are defined as January to March, April to June, July to September, and October to December. The y axis is shown on a logarithmic scale.

were performed by the authors using a shared protocol (rather than through large scale crowd sourced annotation), to support consistent application of the inclusion and exclusion criteria. We restricted candidates to videos that are publicly viewable on YouTube and accessed through features of the service in accordance with the YouTube Terms of Service and the uploader selected YouTube licence type, noting that YouTube supports the Standard YouTube licence by default and the Creative Commons Attribution licence where specified [4]. Then each candidate upload was manually selected and only continuous driving portions that met the eligibility criteria below were retained. Screening consisted of a direct visual review of the footage to verify compliance and determine segment boundaries, as follows:

- **Target duration:** each retained segment is 5 minutes long.
- **Predominantly urban:** segments primarily depict streets and junctions within built up urban areas. Short intervals that pass through urban parks or green spaces within a locality are allowed, but segments dominated by motorways, rural roads, large parking areas, or other clearly nonurban contexts are excluded.
- **Routine conditions:** segments dominated by atypical events or unusual disruptions are excluded, for example, collisions and their aftermath, near misses, confrontations, police stops, emergency response scenes, parades, protests, festivals, or other organised events that materially alter the driving context. We note that it is not always possible to determine whether a particular day is representative from video alone. The collection period overlaps with the Covid 19 era, and some segments may reflect pandemic related traffic and behaviour changes.
- **Continuous unedited dashcam view:** segments containing edited discontinuities or non dashcam intervals are excluded. Edited discontinuities are defined as explicit edits that remove intervening time

[4] `https://support.google.com/youtube/answer/9783148?hl=en&sjid=15816069292466875886-EU`

or change context, including hard cuts or jump cuts, cross fades or other transitions, inserted title cards or chapter separators that skip to a different time or location, time lapse or accelerated playback, and montage or compilation structures.

- **On-screen text and overlays:** videos sometimes contain captions, watermarks, channel branding, timestamps, or other labels. These are permitted when they do not substantially occlude the driving scene. Candidates with persistent overlays that occupy a substantial portion of the frame are excluded.
- **Stationary non-traffic stops:** intervals where the vehicle is stationary for reasons unrelated to normal traffic flow are excluded, including prolonged stops in car parks, petrol stations, service areas, drive through queues, and similar situations. This criterion is applied on the basis of the scene context and the duration of the stationary interval.

When a video contained multiple compliant driving portions separated by excluded content, we retained each continuous compliant portion as a separate segment and recorded its start and end times (in seconds) relative to the beginning of the original upload. Searches and screening were conducted by the authors from 3 February 2024 to 23 February 2026.

Table 2: Distribution of retained CROWD segments and total retained duration by continent, with counts of unique countries and localities. *Segment share* is computed over all retained segment records; *duration share* is computed over total retained hours. Continents follow the UN Statistics Division (UNSD) M49 geographic regions convention[55].[5]

Continent	Countries	Localities	Segment records	Segment share (%)	Duration share (%)	Duration (h)
Europe	53	3,167	19,989	38.62	36.08	7,315.56
Asia	52	1,552	13,707	26.49	26.73	5,418.74
North America	37	1,345	10,851	20.97	25.61	5,192.71
Africa	59	516	2,825	5.46	3.33	675.88
Oceania	22	365	2,468	4.77	4.25	861.50
South America	18	158	1,913	3.70	4.00	811.17
Total	238	7,103	51,753	100.00	100.00	20,275.56

To standardise ingestion and minimise duplicate locality entries, we used a structured curation workflow implemented as a web based form. For each candidate upload, the curator entered a locality name, an optional state or region, the country, and the YouTube URL. The locality names were normalised using a canonical *locality* field and a companion *locality aliases* field, which records alternative spellings as well as historical or local names for the same locality (for example, *Copenhagen*, *Kobenhavn*, *København*). Submitted names were matched against both canonical names and recorded aliases, with country and, when provided, state or region used to disambiguate homonymous localities. When a matching locality record already existed, new uploads were appended to that record; otherwise, a new locality record was created. We record aliases primarily in Latin script, and alternative names written only in non Latin scripts are not systematically included in the *locality aliases* field.

For each submitted URL, the system resolved the YouTube upload identifier and retrieved basic platform metadata, including the channel identifier and publish date. The curator then specified segment start and end times in seconds relative to the beginning of the upload and assigned manual labels for time of day and recording platform type. Validation checks enforced that the end time exceeded the start time and that categorical fields took values from the permitted label sets.

The time of day was assigned by visual inspection using street lighting as an operational cue. Each retained segment was assigned exactly one time-of-day label: a segment was labelled night when street lights were illuminated and remained consistently visible; otherwise it was labelled day. When a single upload

[5]In the underlying mapping, 4 uploads are associated with more than one continent. The segment and duration totals in Table 2 are computed by summing over mapping rows within each continent, so these uploads contribute to every continent to which they are mapped. For counts of unique uploads by continent, each upload is assigned to a single continent: we first take the most frequent continent among its mapped rows; if there is a tie, we assign the upload to the continent with the greatest total mapped duration (in seconds) for that upload.

contained both daylight and nighttime driving within retained footage, the retained portion was split into separate segments at the onset of consistent street-light illumination so that each segment had a single label.

The recording platform was annotated with the following base classes: *Car*, *Bus*, *Truck*, *Two wheeler*, *Bicycle*, *Electric scooter*, and *Unicycle*. In addition, an automation status was recorded when applicable, yielding automated variants of the relevant classes (for example, *Automated car*, *Automated bus*, *Automated truck*, and *Automated two wheeler*). When platform type was difficult to infer from the forward facing view, the authors used additional contextual cues, including the vehicle silhouette, shadow, or reflection (when visible, for example, in windows or in the bodywork of nearby vehicles), apparent camera height and motion, and the width and vertical clearance of the path being followed (for example, the clearance and lane use typically required by a car compared to a bicycle). Automated variants were assigned based on the video title, description, and, where available, chapter information.

To improve the consistency of manual labels and segment boundary definitions, we incorporated an independent audit step during curation. One author periodically reviewed a random sample of segments added in the preceding two weeks by the other authors. For each sampled segment, the auditor re-watched the footage and verified segment start and end times, the time-of-day label, and recording platform type. Any discrepancies were corrected in the curation database, and ambiguous cases were resolved by discussion among the authors. This procedure was intended as pragmatic quality control during dataset construction rather than a formal inter-annotator agreement study.

2.3 Video retrieval and processing

Videos were downloaded from YouTube as MP4 files. The pipeline first attempted retrieval with *pytubefix* (`https://pytubefix.readthedocs.io/en/latest/`) and if that failed, retried with *yt_dlp* (`https://pypi.org/project/yt-dlp/`). In both cases, a single stream was selected by resolution. The downloader first searched for an exact match among the preferred resolutions of 720p, 480p, 360p, and 144p. When no exact match was available, it selected the highest available MP4 stream with frame height at most 720 pixels, preferring a progressive stream when multiple streams shared the same height. If no MP4 stream at or below 720 pixels was available, it selected the lowest available MP4 stream above 720 pixels, again preferring a progressive stream when possible. In the *yt_dlp* fallback, only non HLS MP4 video formats were considered. After download, the frame rate was obtained from the saved video file using OpenCV (`https://opencv.org`) and rounded to the nearest integer. This rounded FPS value was then used in the output file names and in subsequent segment processing and tracking.

Each retained upload was trimmed into the manually curated segments. To reduce boundary artefacts at segment endpoints, we used an adjusted segment end time of ($t_{\text{end}} = 1$) seconds. All retained segments were processed with the You Only Look Once object detector [56] using Ultralytics YOLOv11x weights (`yolo11x.pt`) trained on MS COCO [57]. For each frame, we recorded object detections as bounding boxes with class labels and confidence scores. The tracking was run through `YOLO(model).track` with `persist=True`, `conf=0.0`, `save=False` and `device=cuda` when available (otherwise, CPU). Parameters not explicitly set in this call (for example, input image size, NMS IoU threshold, maximum detections) followed the Ultralytics defaults for the installed `ultralytics` version (see section 7).

Detections were linked across frames using the BoT-SORT tracker [58] with the hyperparameters in Table 8. We enabled re-identification (`with_reid=True`) and global motion compensation (`sparseOptFlow`). The track buffer was specified in seconds (`track_buffer_sec=2`) and converted to frames per segment by setting `track_buffer = track_buffer_sec * FPS`. Tracking was reinitialised at each segment boundary; therefore, track identifiers are unique only within a segment.

2.4 Contextual indicators and metadata

To support cross-locality comparisons, each locality record was enriched with contextual indicators retrieved at data entry time using the submitted locality, optional state/region, and country. Geographic coordinates (latitude and longitude) were recorded to support visualisation and location-linked queries. When available, coordinates were taken directly from video metadata; when coordinates could not be retrieved, they were obtained via manual online lookup using a standardised location query (*locality, state/region (if available), country*) and entered during curation.

Table 3: Upload-level time-of-day composition within each continent. An upload is *day only* if all retained segments are labelled day, *night only* if all are labelled night, and *both* if it includes at least one of each. Uploads are assigned a canonical continent (mode over linked segments) so that per-continent totals sum to the global unique-upload total. Percentages are computed within continent.

Continent	Day	Night	Both day & night	Unique uploads
Europe	14,258 (87.33%)	1,545 (9.46%)	523 (3.20%)	16,326
Asia	8,960 (81.48%)	1,605 (14.60%)	431 (3.92%)	10,996
North America	7,926 (85.51%)	918 (9.90%)	425 (4.59%)	9,269
Africa	1,779 (93.73%)	79 (4.16%)	40 (2.11%)	1,898
Oceania	1,729 (89.68%)	147 (7.62%)	52 (2.70%)	1,928
South America	1,262 (78.14%)	278 (17.21%)	75 (4.64%)	1,615
Total	35,914 (85.44%)	4,572 (10.88%)	1,546 (3.68%)	42,032

The resulting locality record includes population, road traffic mortality, income inequality, gross metropolitan product (when available), literacy rate, and a traffic index, along with geographic coordinates and a continent assignment based solely on geographic location. National level indicators are assigned using a sovereign country mapping, which can differ from the local territory name in the `country` field. For example, Cayenne is geographically in South America, but national indicators are taken from France; similarly, Spanish enclaves in North Africa are geographically assigned to Africa, while national indicators are taken from Spain. Source services used in the curation pipeline include REST Countries (`https://restcountries.com/`) for country attributes (for example, population, ISO codes, and Gini), the World Bank indicators for traffic mortality (`https://data.worldbank.org/indicator/SH.STA.TRAF.P5`) and literacy (`https://data.worldbank.org/indicator/SE.ADT.LITR.ZS`), and Numbeo for a location-based traffic index (`https://www.numbeo.com/traffic/rankings.jsp`). Auxiliary tables were used to provide summary demographic statistics such as median age and average height.

For each video and segment, we also captured YouTube-provided metadata including title, upload date, channel, view count, description, and, where available, chapter information, alongside segment definitions and the derived detection and tracking outputs.

Table 4: Global unique-video counts by vehicle type. Percentages are computed relative to the global unique-upload total ($n = 42,032$).

Vehicle type	Videos	Share (%)
Automated car	3	0.01
Bicycle	566	1.35
Bus	926	2.20
Car	39,278	93.45
Electric scooter	43	0.10
Monowheel/unicycle	38	0.09
Truck	342	0.81
Two-wheeler	845	2.01

2.5 Ethics statement

This dataset is derived from publicly available dashcam videos hosted on *YouTube* and accessed for research purposes according to the platform terms. The release does not redistribute any underlying video files; it provides only YouTube upload identifiers, segment timestamps, and derived annotations (object detection and track identifiers). Because the source videos may contain identifiable individuals and vehicles (e.g. faces and licence plates), we treat the underlying content as potentially sensitive. The annotations released are

limited to bounding boxes, class labels, confidence scores, and segment-local track identifiers; they do not include identity labels, biometric templates, or any attempt to infer sensitive personal attributes.

Users who retrieve and process the referenced videos are responsible for ensuring compliance with applicable laws and regulations (including privacy and data-protection requirements) and with platform terms and copyright restrictions. We encourage downstream users to avoid attempts to identify individuals and to apply appropriate safeguards if working with subsets that may contain sensitive scenes. Some referenced uploads may become unavailable over time due to removal, restriction, or account changes; the dataset should be treated as an index to third-party content rather than an archival copy. When using this resource, we ask users to cite this Data Descriptor and the primary sources for any third party tools used to generate derived annotations. If access to specific referenced uploads is not possible, or if there are questions about use beyond what is enabled by the hosting platform, users are welcome to contact the authors for guidance.

Table 5: Schema of `mapping.csv`. List valued fields are stored as text using square bracket notation, for example `[a,b,c]`. Indices align with `videos`. Nested lists in `time_of_day`, `start_time`, and `end_time` represent multiple retained segments within one upload.

Field	Type	Units	Notes
`id`	int	–	Locality record identifier, unique per row, > 0.
`locality`	string	–	Canonical locality name.
`locality_aka`	string	–	Locality aliases stored as a text encoded list (for example `["Kobenhavn","København"]`).
`state`	string	–	Optional state or region, may be empty.
`country`	string	–	Country or territory name.
`iso3`	string	–	ISO3 code.
`continent`	string	–	Geographic continent from coordinates.
`lat`	float	degrees	Latitude in $[-90, 90]$.
`lon`	float	degrees	Longitude in $[-180, 180]$.
`gmp`	float	–	Gross metropolitan product when available, 0.0 if unavailable.
`population_locality`	int	persons	Locality population estimate, ≥ 0.
`population_country`	int	persons	Country or territory population, ≥ 0.
`traffic_mortality`	float	–	Road traffic mortality rate.
`literacy_rate`	float	%	Literacy rate, in $[0, 100]$.
`avg_height`	float	cm	National average height, ≥ 0.
`med_age`	float	years	Median age, ≥ 0.
`gini`	float	–	Income inequality indicator, typically in $[0, 100]$.
`traffic_index`	float	–	Location based traffic index, ≥ 0.
`videos`	string	–	List of YouTube video IDs stored as text, for example `[QZPqluJAOOY,...]`.
`time_of_day`	string	–	Nested lists stored as text. Outer indices align with `videos`. Inner lists give per segment labels, 0 means day and 1 means night.
`start_time`	string	seconds	Nested lists stored as text. Segment start times in seconds from upload start. Outer indices align with `videos`.
`end_time`	string	seconds	Nested lists stored as text. Segment end times in seconds from upload start. Outer indices align with `videos`.
`vehicle_type`	string	–	List stored as text with one platform label per video, aligned with `videos`.
`upload_date`	string	–	List stored as text with one date per video, aligned with `videos`, stored as `DDMMYYYY`[6].
`channel`	string	–	List stored as text with one channel identifier per video, aligned with `videos`.

[6]Some YouTube metadata values are missing when an upload becomes unavailable (for example made private or removed) before platform metadata could be retrieved. Missing entries are left empty and remain aligned by index with `videos`. This can affect `upload_date` (and `channel`) in `mapping.csv`, and `title`, `upload_date`, `channel`, `description`, and `chapters` in `mapping_metadata.csv`.

3 Data records

The CROWD dataset is available through 4TU.ResearchData [59]. The deposited release comprises two parts: (i) two index tables in CSV format, `mapping.csv` and `mapping_metadata.csv`; and (ii) CSV files per segment containing per frame bounding box annotations from YOLOv11x linked to BoT SORT tracks. The `mapping.csv` file has one row per locality (with an optional state or region for disambiguation). It stores locality descriptors and indicators, including `locality`, `locality_aka`, `state`, `country`, `iso3`, `continent`, `lat`, `lon`, and contextual fields such as population and traffic statistics. Each locality row also aggregates the related YouTube videos in the list valued field `videos`. Segment information is stored in nested lists aligned with `videos`: `time_of_day`, `start_time`, and `end_time` are lists of lists, where the outer index matches the corresponding entry in `videos`, and each inner list contains one entry per retained segment for that video. Fields such as `vehicle_type`, `upload_date`, and `channel` are stored as lists with one entry per video, aligned by the index to `videos`.

The `mapping_metadata.csv` file has one row per YouTube video and contains the YouTube video identifier in the `video` field, together with `title`, `upload_date`, `channel`, `views`, `description`, `chapters` where available, `segments` and `date_updated`. The two index tables are linked through the YouTube video identifier, that is, the entries in `mapping.csv` `videos` correspond to the entries in `mapping_metadata.csv` `video`. Each segment annotation file is named `{video_id}_{start_time}_{fps}.csv` and contains per frame bounding box detections, confidence scores, a segment local track identifier, and the frame index. The schema definitions for `mapping.csv`, `mapping_metadata.csv`, and the annotation files are provided in Table 5, Table 6, and Table 7.

The mapping table is organised around canonical locality records, optionally disambiguated by state or region, and aggregates one or more uploads and their retained segments for each locality. Each row stores locality coordinates and contextual indicators, as well as aligned lists describing associated uploads, segment boundaries (in seconds) relative to each upload, time-of-day labels, recording platform-type labels, and basic YouTube metadata (including upload date and channel). List-valued fields are aligned by index to the `videos` list. The nested list fields (`time_of_day`, `start_time`, and `end_time`) store multiple retained segments per upload, where each inner list corresponds to the upload at the same index in `videos`, and elements within each inner list align by position across the three fields (see Table 5). The *locality aliases* field is included to preserve a consistent locality identity across alternative spellings and local or historical names during data entry, as described in section 2. To support interpretation of the mapping table, we provide summary visualisations of coverage, density, and temporal provenance. Figure 1 shows the geographic distribution of localities with at least one retained segment. Figure 2 reports the upload-month distribution of the underlying YouTube videos, derived from the upload date metadata provided by the platform.

For each retained segment, we provide per-frame detection and tracking output in CSV format. The field `yolo-id` corresponds to the COCO class index emitted by the Ultralytics YOLO model weights (typically 0–79 for COCO-trained models), bounding boxes are provided in YOLO format as centre coordinates and dimensions normalised to $[0, 1]$ relative to the frame extent, and `unique-id` is the BoT-SORT track identifier, which is unique within each segment across all retained classes and is not comparable between segments. The frame indices (`frame-count`) start at 1 for the first frame of the segment. A detection timestamp within the source upload can be recovered as `start_time + (frame-count-1)/fps`. The BoT-SORT configuration used to generate the released track identifiers is documented in Table 8.

Videos are attributed to continents through the locality associated with each retained segment. Because a single YouTube upload can contribute more than one segment, continental coverage is reported at the segment level (see Table 2). Time of day labels are defined at the segment level and summarised at the upload level. An upload is labelled *both day and night* if it has at least one retained segment labelled day and at least one retained segment labelled night, for example, when a single upload is split into separate day and night segments (see Table 3). The distribution of unique videos by vehicle type is shown in Table 4. Videos labelled as car recordings account for the majority of the dataset, whereas buses, trucks, bicycles, two wheelers, and automated and micro mobility classes occur less frequently.

Table 6: Schema of `mapping_metadata.csv`. This table has one row per YouTube video. The `chapters` field is stored as text using square bracket notation and is not JSON. The YouTube video identifier in `video` can be used to link to entries in the `videos` field of `mapping.csv`.

Field	Type	Units	Notes
`id`	int	–	Video record identifier, > 0.
`video`	string	–	YouTube video identifier used for linking, for example `ID_q1RC_dSo`.
`title`	string	–	Video title text, may be empty if unavailable.
`upload_date`	int	–	Upload date stored as `DDMMYYYY`, may be empty if unavailable.
`channel`	string	–	Channel identifier, may be empty if unavailable.
`views`	int	count	View count at time of collection, ≥ 0.
`description`	string	–	Video description text, may be empty if unavailable.
`chapters`	string	–	Chapter list stored as text, either `[]` or a list of chapter records with `title` and `timestamp`, for example `['title': 'START', 'timestamp': '0:00:00', ...]`.
`segments`	int	count	Number of retained segments linked to this video, ≥ 0.
`date_updated`	int	–	Date this record was last updated, stored as `DDMMYYYY`.

4 Technical validation

To ensure internal consistency and release integrity, we implemented an automated release *validator* that operates directly on the distributed CSV artefacts, with no intermediate conversion steps. The validator produces three reproducibility artefacts: (i) a machine readable validation report, (ii) a human readable execution log, and (iii) a cryptographic checksum manifest for the full release. This validation focuses on the structural correctness, completeness, and coherence between the released artefacts, and it does not constitute a manual assessment of detection or tracking accuracy.

Table 7: Schema of per segment YOLO and BoT SORT annotation files named {`video_id`}_{`start_time`}_{`fps`}`.csv`. Column names follow the CSV header.

Field	Type	Units	Notes
`yolo-id`	int	–	COCO class index, from 0 to 79.
`x-center`	float	–	Box centre x, normalised by image width, in $[0, 1]$.
`y-center`	float	–	Box centre y, normalised by image height, in $[0, 1]$.
`width`	float	–	Box width, normalised by image width, in $[0, 1]$.
`height`	float	–	Box height, normalised by image height, in $[0, 1]$.
`unique-id`	int	–	BoT SORT track identifier, unique within the segment, ≥ 0.
`confidence`	float	–	Detector confidence score in $[0, 1]$.
`frame-count`	int	frames	Frame index within the segment, first frame is 1, ≥ 1.

Internal consistency of the mapping table. The validator parses the mapping table and verifies that the list encoded fields are structurally aligned. For each mapping row, it checks that the outer list lengths describing videos and their associated attributes, for example time of day labels and segment boundary lists, are consistent. Within each video entry, it validates segment boundaries by requiring finite numeric timestamps with non negative values and strictly ordered intervals, with start $<$ end. These checks detect mis defined rows, misaligned list encodings, and invalid segment annotations.

Recomputation of paper aligned dataset summaries. Using the validated mapping content, the validator recomputes the primary descriptive statistics reported in the manuscript: the number of unique videos, the number of upload records, the number of segment records, and the total annotated duration obtained by summing end minus start across segments. It also reproduces continent level aggregations of

segment records and duration, continent wise day and night label entry distributions, and global as well as continent stratified upload composition, day only, night only, both, and unknown. Where reference totals are available, the recomputed values are compared with those references and any discrepancies are flagged.

Cross file consistency between mapping and per video metadata. The validator performs bidirectional consistency checks between the mapping table and the per video metadata table. It verifies that every video referenced in the mapping table has a corresponding metadata record and reports any metadata videos that are absent from the mapping, excluding placeholder identifiers. In addition, it checks that per video segment counts stored in the metadata are consistent with the number of unique segments derived from the mapping table for each video, after deduplicating segments by their start and end boundaries. These checks ensure that per video summaries are coherent across release components and that no videos are missing in either direction.

Integrity and coverage of detection outputs. The validator scans the folder of detector produced CSV files and enforces a naming convention that encodes the source video identifier, segment start time, and frame rate. For each detection file, it verifies that the referenced video identifier exists in the mapping table and that the encoded segment start time corresponds, within a fixed tolerance, to a known segment start time for that video derived from the mapping table. The validator detects duplicate associations, where multiple files are assigned to the same segment, and computes segment coverage as the fraction of segments derived from the mapping for which a detection file is present.

Reproducibility through a cryptographic manifest. Finally, the validator computes SHA 256 hashes for all released artefacts and records them in a checksum manifest. Manifest generation is implemented to be robust to absolute paths when artefacts are stored outside the repository root, whilst still emitting stable relative paths where possible. This manifest enables third parties to verify file integrity and detect unintended changes across releases. Because detector and tracker outputs may be non deterministic across environments or when processing begins from different temporal offsets, the release should be treated as the output of a canonical processing run, and the checksum manifest allows users to confirm they have obtained the exact distributed artefacts.

5 Usage Notes

The dataset distributes YouTube upload identifiers, segment start and end times, and derived annotations, but does not redistribute the underlying video files. Users can access the referenced videos through YouTube using the provided identifiers. For analyses that require local video processing, we provide the code used in this work to retrieve and preprocess the referenced uploads in Section 7; use of this code is subject to platform terms and video availability. Because the referenced videos are hosted by YouTube, some uploads may become unavailable over time due to removal, restriction, regional limitations, or account changes.

Users should be aware of several limitations relevant to reuse. The object detections and tracks are derived from a detector trained on the MS COCO label set and therefore cover only the corresponding classes (e.g. persons, bicycles, motorcycles, cars, buses, trucks, traffic lights, and stop signs). The annotations do not include lane markings, road geometry, or a comprehensive traffic sign taxonomy beyond the COCO categories. Tracking performance may degrade under occlusion, dense traffic, motion blur, low illumination, or adverse weather, and track identifiers are not propagated across segment boundaries because tracking is reinitialised per segment.

The coverage of time-of-day is uneven: most retained content is driving during the day (Table 3). Analyses or models intended for nighttime conditions should therefore account for this imbalance (for example, by stratified sampling, weighting, or separate evaluation by time-of-day label).

Geographic coverage is also uneven. Africa is the second largest continent by land area and the second most populous continent after Asia, yet it contributes the smallest retained duration in our collection (Table 2). Studies that aim to make claims across regions should therefore consider continent specific sampling strategies and report results stratified by continent and locality.

Table 8: BoT-SORT tracking hyperparameters used to generate segment level track identifiers in the released per segment CSV files.

Hyperparameter	Value	Description
track high thresh	0.7	Confidence threshold for the first association stage during tracking.
track low thresh	0.3	Confidence threshold for the second association stage during tracking.
new track thresh	0.7	Minimum confidence required to initialise a new track when no matches are found.
track buffer time	2	Duration in seconds to keep lost tracks alive before removal. Converted per segment to `track_buffer` in frames as $2 \times$ FPS.
matching thresh	0.6	Similarity threshold used to match existing tracks with detections.
use fused score	True	If enabled, combines detection confidence with IoU distance for matching.
gmc method	sparseOptFlow	Global motion compensation method used for moving camera footage.
proximity thresh	0.5	IoU threshold for spatial proximity during association.
appearance thresh	0.25	Threshold for appearance based matching.
use reid	True	Enables appearance based re identification.
reid model	auto	Re-identification model selection used by the tracker. The value auto uses the Ultralytics default.

6 Data Availability

The CROWD dataset is available via 4TU.ResearchData at `https://doi.org/10.4121/06e9bb9a-a064-412b-b0f3-9ac5dd62ea16`. The deposited release contains the mapping and metadata index tables, the per-segment derived annotation CSV files, and accompanying documentation.

The dataset release is distributed under the *Creative Commons Attribution 4.0 International (CC BY 4.0)* licence. This licence applies only to the materials included in the deposited release and does not apply to the underlying third-party video content referenced by YouTube identifiers, which remains subject to the original platform terms and copyright restrictions.

Users should cite the specific released dataset version used in their study. Because referenced uploads may become unavailable over time and index files may be updated accordingly, version specific citation supports reproducibility.

7 Code Availability

The source code to download videos, curate segments and generate the released per-segment annotation CSVs is available at `https://github.com/Shaadalam9/pedestrians-in-youtube`. The repository contains the exact configuration files used for the release (e.g. `config`, `bbox_custom_tracker.yaml`), and an environment specification (`pyproject.toml`) targeting Python 3.10.18 with PyTorch 2.8.0 and `ultralytics` v8.3.182 or later.

References

[1] World Health Organization. *Global status report on road safety 2023*. World Health Organization, Geneva, 2023. ISBN 978-92-4-008651-7. URL `https://www.who.int/publications/i/item/9789240086517`. Global report.

[2] Department for Transport, UK Government. Road safety open data, 2025. URL https://www.gov.uk/government/statistical-data-sets/road-safety-open-data.

[3] National Highway Traffic Safety Administration (NHTSA), U.S. Department of Transportation. Fatality analysis reporting system (fars), 2025. URL https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.

[4] T. A. Dingus, S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. R. Knipling. The 100-car naturalistic driving study, phase ii: Results of the 100-car field experiment. Technical report (dot hs 810 593), National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, DC, April 2006. URL https://www.nhtsa.gov/sites/nhtsa.gov/files/100carmain.pdf.

[5] Jonathan F Antin, Suzie Lee, Miguel A Perez, Thomas A Dingus, Jonathan M Hankey, and Ann Brach. Second strategic highway research program naturalistic driving study methods. *Safety Science*, 119: 2–10, 2019. doi: https://doi.org/10.1016/j.ssci.2019.01.016.

[6] Rob Eenink, Yvonne Barnard, Martin Baumann, Xavier Augros, and Fabian Utesch. Udrive: the european naturalistic driving study. In *Proceedings of Transport Research Arena*. IFSTTAR, 2014. doi: 10.1007/s12544-016-0202-z.

[7] Transportation Research Board. Safety data access — strategic highway research program 2 (shrp 2). Transportation Research Board (TRB), National Academies of Sciences, Engineering, and Medicine, 2010. URL https://www.trb.org/StrategicHighwayResearchProgram2SHRP2/SHRP2DataSafetyAccess.aspx. Accessed: 2026-02-22.

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. doi: https://doi.org/10.1109/CVPR.2012.6248074.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. doi: https://doi.org/10.1109/CVPR.2016.350.

[10] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018. doi: https://doi.org/10.1109/CVPRW.2018.00141.

[11] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. doi: https://doi.org/10.1109/cvpr42600.2020.00271.

[12] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. doi: 10.1109/ICCV.2017.534.

[13] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3292–3310, 2022. doi: https://doi.org/10.1109/TPAMI.2022.3179507.

[14] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. doi: https://doi.org/10.48550/arXiv.2004.06320.

[15] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. doi: https://doi.org/10.1109/CVPR.2019.00895.

[16] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. doi: https://doi.org/10.48550/arXiv.2301.00493.

[17] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. doi: https://doi.org/10.1109/CVPR42600.2020.01164.

[18] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. doi: https://doi.org/10.1109/CVPR42600.2020.00252.

[19] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pages 3095–3101. IEEE, 2021. doi: https://doi.org/10.1109/ITSC48978.2021.9565009.

[20] Markus Braun, Sebastian Krebs, Fabian B. Flohr, and Dariu M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. ISSN 0162-8828. doi: https://doi.org/10.1109/TPAMI.2019.2897684.

[21] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6262–6271, 2019. doi: https://doi.org/10.1109/ICCV.2019.00636.

[22] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*, 2016. doi: https://doi.org/10.48550/arXiv.1609.04741.

[23] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. doi: https://doi.org/10.1109/CVPR.2011.5995347.

[24] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. Dada-2000: Can driving accident be predicted by driver attentionf analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 4303–4309. IEEE, 2019. doi: https://doi.org/10.1109/ITSC.2019.8917218.

[25] International Organization for Standardization. ISO 3166 — country codes. `https://www.iso.org/iso-3166-country-codes.html`. Accessed 3 March 2026.

[26] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern recognition letters*, 30(2):88–97, 2009.

[27] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, pages 304–311. IEEE, 2009.

[28] Sangkeun Park, Joohyun Kim, Rabeb Mizouni, and Uichin Lee. Motives and concerns of dashcam video sharing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4758–4769, 2016.

[29] Joohyun Kim, Sangkeun Park, and Uichin Lee. Dashcam witness: Video sharing motives and privacy concerns across different nations. *IEEE Access*, 8:110425–110437, 2020.

[30] Md Shadab Alam, Marieke H. Martens, and Pavlo Bazilinskyy. Pedestrian planet: What youtube driving from 233 countries and territories teaches us about the world. In *Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '25, page 180–197, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720130. doi: 10.1145/3744333.3747827.

[31] Zhengping Che, Guangyu Li, Tracy Li, Bo Jiang, Xuefeng Shi, Xinsheng Zhang, Ying Lu, Guobin Wu, Yan Liu, and Jieping Ye. D ^{2}-city: a large-scale dashcam video dataset of diverse traffic scenarios. *arXiv preprint arXiv:1904.01975*, 2019. doi: https://doi.org/10.48550/arXiv.1904.01975.

[32] Patrick Wenzel, Rui Wang, Nan Yang, Qing Cheng, Qadeer Khan, Lukas Von Stumberg, Niclas Zeller, and Daniel Cremers. 4seasons: A cross-season dataset for multi-weather slam in autonomous driving. In *DAGM German Conference on Pattern Recognition*, pages 404–417. Springer, 2020.

[33] Michela Cantarini, Leonardo Gabrielli, Adriano Mancini, Stefano Squartini, and Roberto Longo. A3carscene: An audio-visual dataset for driving scene understanding. *Data in Brief*, 48:109146, 2023. doi: https://doi.org/10.1016/j.dib.2023.109146.

[34] Keenan Burnett, David J Yoon, Yuchen Wu, Andrew Z Li, Haowei Zhang, Shichen Lu, Jingxing Qian, Wei-Kang Tseng, Andrew Lambert, Keith YK Leung, et al. Boreas: A multi-season autonomous driving dataset. *The International Journal of Robotics Research*, 42(1-2):33–42, 2023.

[35] Matthew Pitropov, Danson Evan Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5):681–690, 2021. doi: https://doi.org/10.1177/0278364920979368.

[36] Harald Schafer, Eder Santana, Andrew Haden, and Riccardo Biasini. A commute in data: The comma2k19 dataset. *arXiv preprint arXiv:1812.05752*, 2018. doi: https://doi.org/10.48550/arXiv.1812.05752.

[37] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41 (7):1720–1733, 2018. doi: https://doi.org/10.1109/TPAMI.2018.2845370.

[38] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018. doi: https://doi.org/10.1109/CVPR.2018.00803.

[39] Ravi Shankar Mishra, Chirag Parikh, Anbumani Subramanian, CV Jawahar, and Ravi Kiran Sarvadevabhatla. Idd-crs: A comprehensive video dataset for critical road scenarios in unstructured environments. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 309–315. IEEE, 2025. doi: https://doi.org/10.1109/IV64158.2025.11097494.

[40] Chirag Parikh, Rohit Saluja, CV Jawahar, and Ravi Kiran Sarvadevabhatla. Idd-x: A multi-view dataset for ego-relative important object localization and explanation in dense and unstructured traffic. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14815–14821. IEEE, 2024. doi: https://doi.org/10.1109/ICRA57147.2024.10609989.

[41] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. doi: 10.48550/arXiv.2106.11810.

[42] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. doi: https://doi.org/10.48550/arXiv.2106.11037.

[43] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14662–14672, 2024. doi: https://doi.org/10.1109/CVPR52733.2024.01389.

[44] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. doi: https://doi.org/10.1177/0278364916679498.

[45] NVIDIA Corporation. Physicalai-autonomous-vehicles. Hugging Face dataset, 2025. URL `https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles`. Access requires agreeing to the NVIDIA Autonomous Vehicle Dataset License Agreement.

[46] Mina Alibeigi, William Ljungbergh, Adam Tonderski, Georg Hess, Adam Lilja, Carl Lindström, Daria Motorniuk, Junsheng Fu, Jenny Widahl, and Christoffer Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20178–20188, 2023. doi: https://doi.org/10.1109/ICCV51070.2023.01846.

[47] Ahmad Rahimi, Valentin Gerard, Eloi Zablocki, Matthieu Cord, and Alexandre Alahi. Mad: Motion appearance decoupling for efficient driving world models. *arXiv preprint arXiv:2601.09452*, 2026. doi: 10.48550/arXiv.2601.09452.

[48] Sandvine. Global internet phenomena report 2024, 2024. URL `https://www.sandvine.com/hubfs/Sandvine_Redesign_2019/Downloads/2024/GIPR/GIPR%202024.pdf`.

[49] Tobias Lohaus, Sara Yüksekdag, Silja Bellingrath, and Patrizia Thoma. The effects of autonomous sensory meridian response (asmr) videos versus walking tour videos on asmr experience, positive affect and state relaxation. *Plos one*, 18(1):e0277990, 2023. doi: https://doi.org/10.1016/j.jad.2021.12.015.

[50] German Lopez. Asmr, explained: why millions of people are watching youtube videos of someone whispering, 2015. URL `https://www.vox.com/2015/7/15/8965393/asmr-video-youtube-autonomous-sensory-meridian-response`.

[51] Emma L Barratt, Charles Spence, and Nick J Davis. Sensory determinants of the autonomous sensory meridian response (asmr): understanding the triggers. *PeerJ*, 5:e3846, 2017. doi: https://doi.org/10.7717/peerj.3846.

[52] Bo Han. How do youtubers make money? a lesson learned from the most subscribed youtuber channels. *International Journal of Business Information Systems*, 33(1):132–143, 2020. doi: https://doi.org/10.1504/IJBIS.2020.10026504.

[53] Md Shadab Alam and Pavlo Bazilinskyy. Nineteen years of asmr on youtube: A multilingual, theme-level analysis of 42,268 videos. 2026.

[54] United Nations Statistics Division. *Principles and Recommendations for Population and Housing Censuses, Revision 3*. Number Series M No. 67/Rev.3 in Statistical Papers. United Nations, New York, 2017. ISBN 978-92-1-161597-5. URL `https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf`.

[55] United Nations Statistics Division. Standard country or area codes for statistical use (m49). URL `https://unstats.un.org/unsd/methodology/m49/`.

[56] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. doi: 10.48550/arXiv.1506.02640.

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. doi: https://doi.org/10.1007/978-3-319-10602-1_48.

[58] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. doi: https://doi.org/10.48550/arXiv.2206.14651.

[59] Md Shadab Alam, Olena Bazilinska, and Pavlo Bazilinskyy. Crowd: City road observations with dashcams [dataset]. 2026. doi: https://doi.org/10.4121/06e9bb9a-a064-412b-b0f3-9ac5dd62ea16. Version described in this Data Descriptor.

Author Contributions

Conceptualization: Md Shadab Alam, Pavlo Bazilinskyy; Methodology: Md Shadab Alam, Pavlo Bazilinskyy; Software: Md Shadab Alam, Pavlo Bazilinskyy; Data curation: Md Shadab Alam, Olena Bazilinska, Pavlo Bazilinskyy; Investigation: Md Shadab Alam, Pavlo Bazilinskyy; Formal analysis: Md Shadab Alam; Validation: Pavlo Bazilinskyy; Visualization: Md Shadab Alam, Pavlo Bazilinskyy; Writing-original draft: Md Shadab Alam; Writing-review & editing: Olena Bazilinska, Pavlo Bazilinskyy; Supervision: Pavlo Bazilinskyy; Project administration: Pavlo Bazilinskyy; Funding acquisition: Md Shadab Alam, Pavlo Bazilinskyy.
All authors have read and agreed to the published version of the manuscript.

Competing Interests

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

The authors thank Margarida Fortes Ferreira and Aloysia Prakoso for their valuable assistance in sourcing and preparing the video material used in this work. The authors also gratefully acknowledge Linghan Zhang, the Digital Twin Lab (DT Lab) at Eindhoven University of Technology[7], and the TU/e Supercomputing Center (HPC)[8] for providing access to their systems and computational resources used in this study.

Funding

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-1603.

[7] https://www.tue.nl/en/research/institutes/eindhoven-artificial-intelligence-systems-institute/digital-twin-lab

[8] https://supercomputing.tue.nl/