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, variable duration, 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 75,147 segment records spanning 26,273.75 hours (58,467 unique uploads), covering 8,049 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 reproducible baseline analysis, we provide per-segment CSV files of automatic detections for all 80 MS-COCO classes produced with YOLOv11x, together with segment-local multi-object tracks generated with 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 detection and tracking pseudo-labels, enabling reproducible research without redistributing the underlying videos.

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 (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). 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. 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]. 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 within retained compliant intervals, 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 and segment-local tracks aligned to video frames as derived baseline outputs. These outputs are intended as automatic pseudo-labels that can support reproducible baselines and follow on experiments in detection, tracking, and interaction analysis without requiring users to rerun the full detection pipeline; they should not be interpreted as manually validated ground truth.

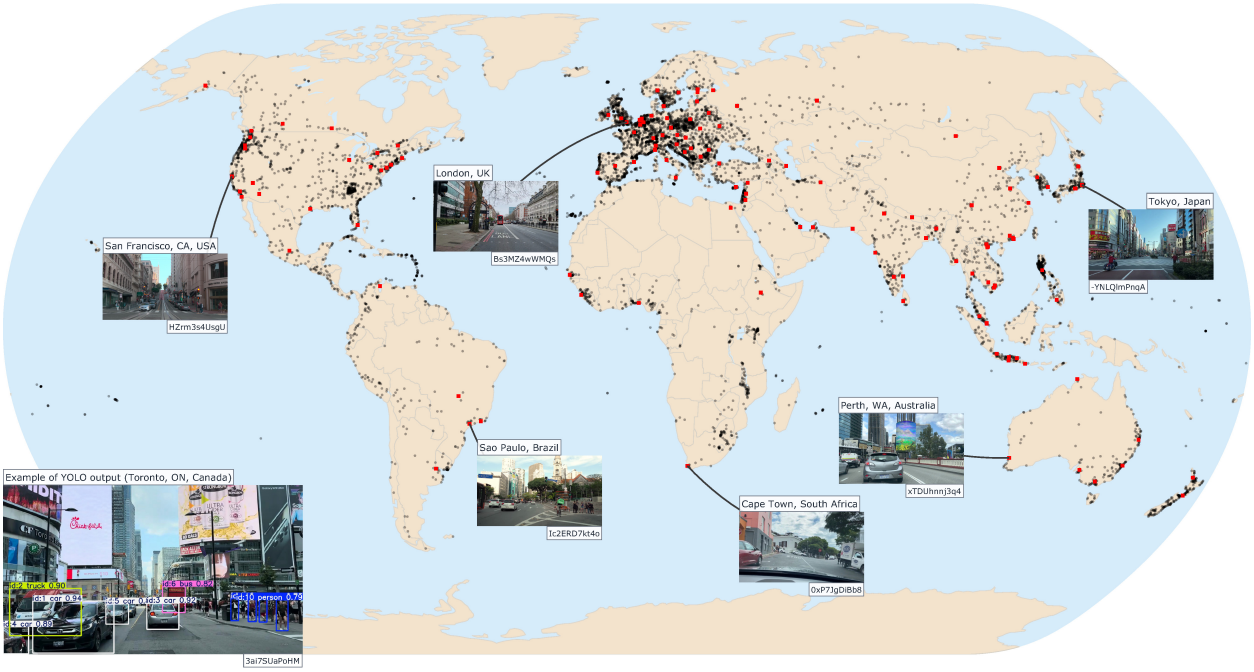

Figure 1: Geographic coverage of CROWD across 8,049 localities worldwide. Each marker denotes one locality record with at least one retained dashcam segment. Marker positions are approximate locality-level coordinates, obtained from the named locality, optional state or region, and country, with manual adjustment where needed during curation. They do not represent route coordinates, GPS traces, exact recording locations, segment start points, or frame-level positions. Red markers indicate localities with more than 100,000 s of retained footage (n = 134); all other localities are shown with the base marker style. Insets show selected example localities linked to their approximate locality positions on the map. Geolocation uncertainty is therefore at the locality scale, and the coordinates should not be interpreted as street-level ground truth.

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 in the notes below the table.

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).

Continued on next page.

Table 1 (continued).

Dataset	Geographic coverage (countries/localities)	Footage (h)	Day/Night	Notes (dashcam video only)
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.
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.

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]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.

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.

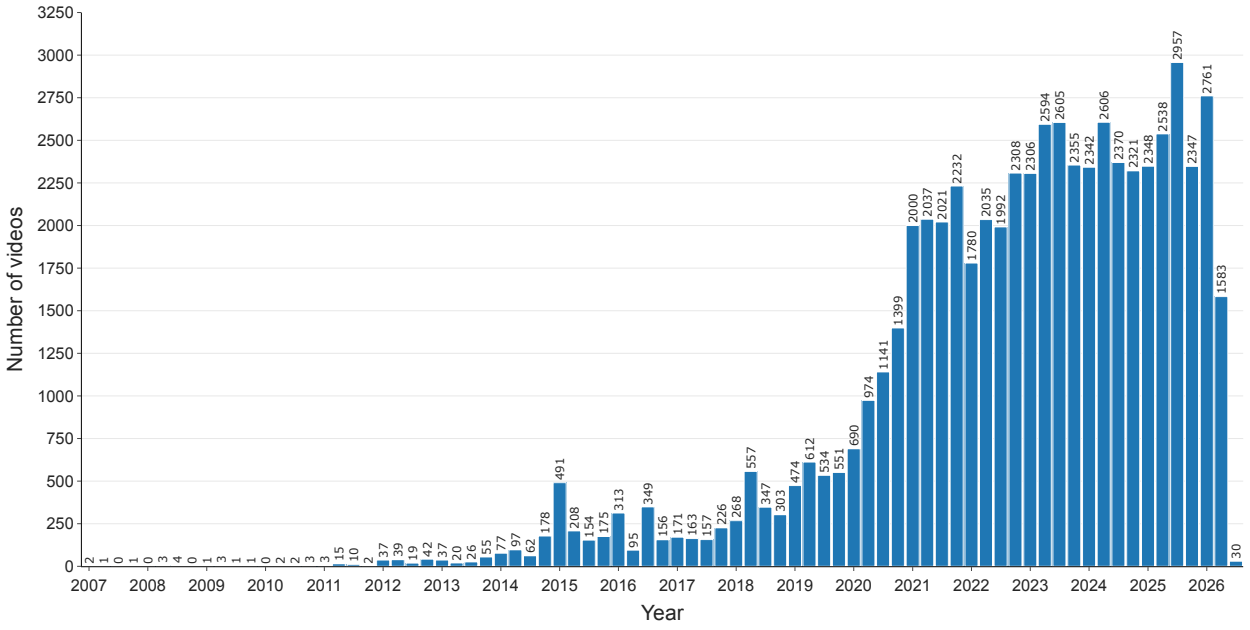

Figure 2: Quarterly platform upload date distribution of YouTube videos with at least one retained CROWD segment. Quarters are January to March, April to June, July to September, and October to December. This distribution reflects YouTube upload dates, not independently verified recording dates, and the date of driving may differ from the date of upload. The 2026 values cover only the incomplete period available at metadata freeze and should not be visually compared with complete calendar years. Metadata for this figure were frozen on 12 July 2026. Six unavailable or invalid upload-date entries were omitted from date parsing.

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.

Methods

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.

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 [country]*, and *dashcam driving in [country]*, where the bracketed terms were replaced by the locality or country names targeted during curation. The input records were publicly viewable YouTube uploads; no videos were obtained from private repositories, paid datasets, organisational collaborations, or non-public data sharing agreements. To allow others to locate the same source records where they remain available, the deposited `mapping.csv` and `mapping_metadata.csv` files include the YouTube upload identifier for each retained video. Each source upload can be resolved using the standard YouTube watch URL pattern `https://www.youtube.com/watch?v=VIDEO_ID`, replacing `VIDEO_ID` with the corresponding identifier in the release. 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 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 (`https://support.google.com/youtube/answer/9783148?hl=en&sjid=15816069292466875886-EU`). 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:

- **Source upload and segment duration:** candidate source uploads were required to be at least five minutes long. This threshold applies to the original YouTube upload, not to the retained segment length. Retained CROWD segments are variable duration compliant intervals, not fixed length five minute windows. Each segment preserves the full continuous portion of an upload that met the eligibility criteria.

- **Predominantly urban:** segments primarily depict streets and junctions within urban areas. This includes ordinary city streets, as well as larger urban roads, such as inner city arterial roads, boulevards, ring roads, and wide multilane roads, when they remain embedded in the urban environment and show urban traffic context such as junctions, pedestrian crossings, sidewalks, roadside buildings, bus stops, parked vehicles, pedestrians, cyclists, or frequent local access. In large car oriented metropolitan areas, some retained footage may also include major urban road corridors within the city, such as parts of Interstate 10 in Los Angeles (CA), where the surrounding scene remains visibly urban. Inclusion is based on the observed roadside context rather than the road name or classification alone. Short intervals are allowed through urban parks or green spaces within a locality. Segments dominated by limited access motorways or freeways, rural roads, large parking areas, or other clearly nonurban contexts are excluded, especially when they lack observable urban roadside activity or vulnerable road user infrastructure.
- **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.
- **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 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.

Table 2: Distribution of retained CROWD segments and total retained duration by continent, with counts of represented countries and territories. *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].

Continent	Countries	Segment records	Segment share (%)	Duration share (%)	Duration (h)
Europe	53	25,202	33.54	32.04	8,419.19
Asia	53	27,057	36.01	33.98	8,927.17
North America	37	12,947	17.23	22.23	5,840.89
Africa	60	2,994	3.98	2.61	684.82
Oceania	22	3,992	5.31	4.99	1,310.27
South America	18	2,955	3.93	4.15	1,091.41
Total	238	75,147	100.00	100.00	26,273.75

Note: In the underlying mapping, 1 upload is 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 this upload contributes to every continent to which it is 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.

When a source upload contained compliant driving footage, each continuous compliant interval was retained as a variable duration segment, with its start and end times recorded in seconds relative to the beginning of the original upload. The segments were defined by compliance boundaries rather than by a

fixed target duration. If excluded content occurred within an otherwise eligible upload, such as a crash scene, non dashcam interval, edited transition, or prolonged stationary non traffic stop, the affected portion was omitted, and the remaining compliant intervals were retained as separate segments where applicable. Ordinary traffic related stopping could remain within a segment. Retained intervals from the same upload were required to be non overlapping, and candidate overlaps identified during curation or release validation were resolved before release. Across the dataset, the 75,147 segment records span 26,273.75 h of processed footage, corresponding to a mean processed segment duration of 20.98 min. Searches and screening were conducted by the authors from 3 February 2024 to 10 July 2026.

Country and locality concentration. To characterise the distribution of retained footage within the geographic coverage reported above, we quantified the retained processing duration by country and by location, using the same endpoint adjustment applied to the duration totals. Table 3(a) reports the ten largest contributing countries by the processing duration retained; together, these countries account for 54.05% of the retained hours. Table 3(b) reports the ten largest contributing localities by retained processed duration. At the locality level, the top 1% of localities (81 localities) contribute 66.32% of retained hours, the top 5% (403 localities) contribute 84.16%, and the top 10% (805 localities) contribute 89.09%. In contrast, 4,451 localities (55.30%) are represented by exactly one segment, and 5,043 localities (62.65%) are represented by exactly one unique upload.

Table 3: Largest contributing countries and localities by retained processed duration. Shares are computed over total retained processed hours.

(a) Countries

Rank	Country	Duration (h)	Share (%)	Segment records	Unique uploads	Localities
1	United States	5,045.85	19.20	10,474	8,855	1,046
2	United Kingdom	1,397.75	5.32	2,737	2,398	330
3	Philippines	1,298.84	4.94	4,076	3,520	400
4	Russia	1,169.37	4.45	2,593	1,877	107
5	Australia	1,133.65	4.31	3,091	2,647	195
6	Vietnam	1,067.49	4.06	3,372	2,774	69
7	Indonesia	968.53	3.69	3,258	2,987	174
8	South Korea	790.17	3.01	1,264	793	83
9	France	710.38	2.70	2,610	1,782	337
10	China	618.61	2.35	1,273	1,088	136

(b) Localities

Rank	Locality	Country	Duration (h)	Share (%)	Segment records	Unique uploads
1	New York, NY	United States	1,276.47	4.86	2,396	2,096
2	London	United Kingdom	1,118.49	4.26	1,655	1,480
3	Los Angeles, CA	United States	852.87	3.25	1,784	1,485
4	Hanoi	Vietnam	810.35	3.08	2,392	1,984
5	Manila	Philippines	776.25	2.95	2,534	2,138
6	Saint Petersburg	Russia	566.03	2.15	771	689
7	Chicago, IL	United States	527.52	2.01	797	666
8	Seoul	South Korea	516.62	1.97	621	259
9	Sydney, NSW	Australia	435.27	1.66	961	860
10	Jakarta	Indonesia	422.70	1.61	1,601	1,532

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 needed for provenance, including the platform upload date where available. The dataset does not include a recording date field because the date of driving could not be independently verified consistently at scale from the source records. Dates mentioned in video titles, descriptions, chapters, or visual overlays were therefore not treated as recording dates. Accordingly, no recording date, recording date provenance, or recording date uncertainty fields are included in this release.

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. At submission, the resolved YouTube upload identifier was checked against all existing mapping records. When the identifier was already present, the interface displayed its existing locality associations and retained segment intervals. A proposed segment was rejected if its interval overlapped any segment already recorded for the same upload, either within the current locality record or elsewhere in the mapping; intervals that shared only an endpoint were permitted. This prevented the same footage interval from being retained more than once while allowing distinct, non-overlapping portions of one upload to be associated with different localities. Additional validation checks enforced that the end time exceeded the start time and that categorical fields took values from the permitted label sets.

Table 4: 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	16,713 (87.84%)	1,740 (9.14%)	574 (3.02%)	19,027
Asia	17,239 (81.15%)	3,200 (15.06%)	805 (3.79%)	21,244
North America	9,251 (86.52%)	996 (9.32%)	445 (4.16%)	10,692
Africa	1,820 (93.77%)	81 (4.17%)	40 (2.06%)	1,941
Oceania	2,844 (90.00%)	239 (7.56%)	77 (2.44%)	3,160
South America	1,927 (80.19%)	384 (15.98%)	92 (3.83%)	2,403
Total	49,794 (85.17%)	6,640 (11.36%)	2,033 (3.48%)	58,467

Each retained segment was assigned a binary contextual time-of-day label by visual inspection. A segment was labelled night when street lamps or comparable artificial road lighting were visibly active and consistently present in the driving scene; otherwise it was labelled day. If retained footage transitioned between daylight and night under this rule, it was split at the transition so that each segment had one label. The field is available for every retained segment and is intended for broad dataset stratification; it does not distinguish dawn, dusk, or twilight, nor does it represent astronomical time or measured illumination.

The recording platform label is also complete under the original curation protocol and describes the apparent recording platform or source asserted platform category available during curation. The represented classes are *Automated car*, *Bicycle*, *Bus*, *Car*, *Electric scooter*, *Emergency vehicle*, *Truck*, *Two wheeler*, and *Unicycle*. 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 with a bicycle. The *Automated car* label was assigned only when source provided title, description, or chapter information indicated automated driving; automation status was not inferred from camera height, shadows, or other visual cues alone. These labels should be treated as apparent or source asserted categories, not as independently verified automated driving status.

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 or accuracy estimation study; accordingly, the released time-of-day and platform fields should be used as coarse contextual metadata rather than benchmark-grade ground truth.

Table 5: Global unique-upload counts by apparent recording platform category. Percentages are computed relative to the global unique-upload total ($n = 58,467$).

Vehicle type	Videos	Share (%)
Automated car	24	0.04
Bicycle	1,351	2.31
Bus	943	1.61
Car	53,882	92.14
Electric scooter	82	0.14
Emergency vehicle	304	0.52
Monowheel/unicycle	38	0.06
Truck	631	1.08
Two-wheeler	1,238	2.12

Note: The platform categories are complete under the original curation protocol and are intended for broad filtering and dataset characterisation. They should not be interpreted as independently verified platform identity. In particular, automated categories indicate apparent or source asserted status available during curation and should not be treated as verified automated-driving ground truth.

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. The released `start_time` and `end_time` fields store the original manually selected segment boundaries in seconds relative to the beginning of the YouTube upload. To avoid boundary artefacts during trimming, the temporary videos used for detection and tracking were generated with the processing endpoint $t_{\text{end,processed}} = t_{\text{end}} - 1$ s, with a safeguard that leaves the endpoint unchanged if the adjusted interval would be non-positive. Duration totals reported here are computed as $t_{\text{end,processed}} - t_{\text{start}}$. Across all 75,147 segments, this gives 94,585,486 s (26,273.75 h) of processed footage after applying the endpoint adjustment. For a segment processed at rounded frame rate f_i, the expected frame-count reduction is approximately f_i frames. Frame indices in the released per-segment CSV files are segment-local: frame 1 corresponds to the first processed frame at the segment start, and frame k corresponds approximately to $t_{\text{start}} + (k-1)/f_i$ seconds in the original upload.

All retained segments were processed with the You Only Look Once (YOLO) object detector [56] using Ultralytics YOLO11x weights (`yolo11x.pt`) trained on MS COCO [57]. The model weights have SHA256

checksum 7bc158aa95c0ebfdd87f70f01653c1131b93e92522dbe15c228bcd742e773a24. For each processed frame, the detector produced bounding boxes, MS COCO class labels, and confidence scores. The pinned software environment and effective inference settings used to generate the released detection and tracking pseudo-labels are reported in Table 13. The machine-readable configuration files are documented with the source code (see the Code Availability section).

Table 6: 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 retrieved through an API where available; otherwise entered manually after online lookup, ≥ 0.
population_country	int	persons	Country or territory population retrieved through an API where available; otherwise entered manually after online lookup, ≥ 0.
traffic_mortality	float	–	Road traffic mortality rate retrieved programmatically where available; otherwise entered manually after online lookup.
literacy_rate	float	%	Literacy rate retrieved programmatically where available; otherwise entered manually after online lookup, in $[0, 100]$.
avg_height	float	cm	National average height, ≥ 0.
med_age	float	years	Median age, ≥ 0.
gini	float	–	Income inequality indicator retrieved programmatically where available; otherwise entered manually after online lookup, typically in $[0, 100]$.
traffic_index	float	–	Location based traffic index retrieved programmatically where available; otherwise entered manually after online lookup, ≥ 0.
videos	string	–	List of YouTube video IDs stored as text, for example [QZPqluJA0OY,...].
time_of_day	string	–	Nested lists stored as text. Outer indices align with videos. Inner lists give one binary contextual label per segment, where 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 apparent recording platform or source asserted platform label per video, aligned with videos; automated status is not independently verified.
upload_date	string	–	List stored as text with one platform upload date per video where retrievable, aligned with videos, stored as DDMMYYYY; this is not a recording date.
channel	string	–	List of platform-assigned channel identifiers aligned with videos where retrievable; these are not channel names, handles, or custom URLs.

Note: Platform metadata could not be retrieved for 393 of the 58,467 unique uploads (0.67%). The corresponding source addresses were not publicly accessible when metadata retrieval was attempted, consistent with uploads hav-

ing been removed, made private, or otherwise becoming unavailable; the specific reason could not be determined consistently. Missing values are left empty and remain aligned by index with `videos`. This can affect `upload_date`, `channel`, and other retained platform metadata fields, but not the retained segment boundaries or manually curated labels. The `channel` field stores a platform-assigned channel identifier, not a channel name, handle, or custom URL.

Detections were linked across frames using the BoT-SORT tracker [58] with the custom YAML configuration reported in Table 12. The configuration file has SHA256 checksum `b0abba37f9e4a80451bf50afbe39bee9b54d48778e0ac988741d1899db9a41d8`. It specifies `tracker_type=botsort`, enables re-identification (`with_reid=true`), and uses `sparseOptFlow` for global motion compensation. The YAML file contains an initial `track_buffer` value of 60 at the 30 fps reference rate. Before each segment is processed, this value is overwritten at runtime by setting `track_buffer = 2 * FPS`, using the rounded frame rate of that segment. The effective buffer is therefore frame-rate dependent and corresponds to approximately two seconds of lost-track retention for every segment. The ReID model setting was `model=auto`; no separate ReID weight file was specified, and resolution of this setting followed the behaviour of the pinned Ultralytics 8.4.46 implementation. Tracking was reinitialised at each segment boundary; therefore, track identifiers are unique only within a segment.

The resulting bounding boxes and track identifiers were generated using pretrained Ultralytics YOLO11x weights trained on MS COCO together with the specified BoT-SORT configuration. The detector was not fine tuned on this dataset, and its object detection and tracking accuracy was not independently evaluated at the object level on the CROWD videos. Performance reported for the pretrained model on its original benchmark data was therefore not assumed to transfer directly to this dataset. The outputs were not manually corrected and are released as detection and tracking pseudo labels rather than object level ground truth. Their accuracy may vary by object class, country, locality, lighting, camera quality, traffic density, occlusion, and weather. Downstream studies requiring ground truth annotations should conduct task specific validation or manual review.

Table 7: Schema of `mapping_metadata.csv`. This table has one row per YouTube upload for which platform metadata could be retrieved. The `video` field links to the `videos` field of `mapping.csv`. The `channel` field stores a platform-assigned channel identifier, not a channel name or handle. Source titles, descriptions, and chapter text are excluded from the public release.

Field	Type	Units	Notes
`id`	int	–	Video record identifier, > 0.
`video`	string	–	YouTube video identifier used for linking, for example `ID_q1RC_dSo`.
`upload_date`	int	–	Platform upload date stored as `DDMMYYYY`, may be empty if unavailable; this is not a recording date.
`channel`	string	–	Platform-assigned channel identifier; not a channel name, handle, or custom URL.
`views`	int	count	View count at metadata retrieval time, ≥ 0 when available.
`segments`	int	count	Number of retained segments linked to this video, ≥ 0.
`date_updated`	int	–	Date this metadata record was last updated, stored as `DDMMYYYY`.

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.

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. Locality population, country or territory population, road traffic mortality, literacy rate, Gini, and the traffic index were first retrieved programmatically through

the relevant API or source service used by the curation pipeline. The locality query used the locality, optional state or region, and country. When a service did not return a usable value for any of these indicators, the curator searched publicly available internet sources and entered the value manually. These indicators should therefore be treated as contextual estimates because their sources and reference years may vary. 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 retained segment, the public release provides the source upload identifier and address, segment boundaries, manual labels, contextual locality metadata, and derived detection and tracking outputs. The platform-assigned channel identifier is retained solely to group uploads originating from the same source, quantify source concentration, detect duplicate source associations, and support provenance audits when individual uploads become unavailable. It is not accompanied by the channel name, handle, custom URL, profile information, or inferred uploader attributes. To minimise the redistribution of potentially identifying or copyrighted source material, free text platform metadata, including titles, descriptions, and chapter text, are not included in the public release. The release also excludes source videos, frames, audio, thumbnails, licence-plate transcriptions, identity labels, biometric descriptors, and other personal-information annotations.

Table 8: Top-level schema of `crowd_index.jsonl`. Each line is one valid JSON object representing one retained segment. Nested objects use standard JSON types rather than Python-style list strings.

Field	Type	Units	Notes
`video_id`	string	–	YouTube upload identifier.
`segment_id`	string	–	Unique identifier derived from the upload ID and segment boundaries.
`youtube_url`	string	–	Resolvable source watch address.
`start_time_s`	number	seconds	Retained segment start relative to the source upload.
`end_time_s`	number	seconds	Retained segment end relative to the source upload.
`upload_date`	string/null	–	Platform upload date where available; not a recording date.
`time_of_day_code`	int	–	Manual segment label: 0 for day and 1 for night.
`time_of_day_name`	string	–	Human-readable name corresponding to the time-of-day code.
`vehicle_type_code`	int	–	Apparent or source-asserted recording-platform category code.
`vehicle_type_name`	string	–	Human-readable name corresponding to the platform code.
`location`	object	–	Nested locality, country, ISO3, continent, latitude, and longitude values.
`metadata`	object	–	Limited structured per-upload provenance metadata copied from `mapping_metadata.csv`; values may be null. Source titles, descriptions, and chapter text are excluded.
`mapping_extra`	object	–	Additional contextual fields copied from the source mapping row.
`mapping_row_number`	int	–	Source row used to construct the segment record.

Ethics statement

CROWD is an index and derived pseudo-label release based on publicly viewable dashcam videos hosted on *YouTube*. The project underwent institutional data protection review before large scale processing. The processing was conducted for scientific research using publicly available source material, in accordance with the applicable institutional data-protection framework and the research safeguards described in GDPR Article 89 [59]. Processing was limited to deriving contextual and object-level information; the footage was

not combined with other datasets containing personal data, and no attempt was made to identify individuals. The underlying videos remain governed by the applicable YouTube terms and uploader-selected licence terms.

The public release does not redistribute source video, frames, audio, thumbnails, channel names, channel handles, custom channel URLs, personal names, or source free text such as titles, descriptions, and chapter text. It retains platform-assigned video and channel identifiers, limited structured provenance metadata, segment timestamps, manual segment labels, contextual locality metadata, and derived YOLO/BoT-SORT detection and tracking pseudo-labels. The channel identifier is retained only to group uploads by source, assess source concentration, detect duplicate source associations, and audit provenance when uploads become unavailable. The derived outputs are limited to generic object classes, bounding boxes, confidence scores, frame indices, and segment-local track identifiers. They do not include licence-plate transcriptions, face or person identities, persistent identity labels, facial embeddings, biometric templates, walking-gait descriptors, or inferred sensitive personal attributes. Track identifiers are arbitrary within-segment object tracks and are not comparable across segments.

Public visibility does not remove privacy or data-protection obligations. Downstream users who retrieve or process the referenced videos are responsible for complying with applicable platform terms, copyright, privacy, data-protection, and research-ethics requirements. CROWD should not be used for identification, surveillance, biometric identification, face recognition, number-plate recognition, law-enforcement decisions, individual profiling, or persistent individual tracking. Users who store, redistribute, publish, or otherwise process source frames or clips should obtain any institutional review, lawful basis, permissions, and data-protection assessment required for their own use case.

Data records

The CROWD dataset is available through 4TU.ResearchData [60]. The deposited release contains three groups of data files: (i) the operational source tables `mapping.csv` and `mapping_metadata.csv`; (ii) the derived segment-level index `crowd_index.jsonl`; and (iii) per-segment CSV files containing frame-level YOLO11x detections linked by BoT-SORT. The source code used to create, process, and validate these files is provided separately as described in the Code Availability section.

The `mapping.csv` and `mapping_metadata.csv` files are the operational tables used by the curation and processing code. The `mapping.csv` file has 8,049 rows, one per canonical locality, with an optional state or region for disambiguation. Its aligned upload lists contain 61,725 locality–upload associations representing 58,467 unique uploads; an upload can contribute to more than one locality record. Each row contains locality descriptors, geographic coordinates, contextual indicators, associated YouTube upload identifiers, segment boundaries, time-of-day labels, apparent recording-platform labels, and platform upload dates and platform-assigned channel identifiers where retrievable. Fields containing information for multiple uploads or segments are stored as aligned text-encoded lists. The schema is provided in Table 6.

The `mapping_metadata.csv` file has 58,074 rows and is linked to `mapping.csv` through the YouTube video identifier. It contains limited structured provenance fields for 99.33% of the 58,467 unique uploads, including the platform upload date, platform-assigned channel identifier, view count, number of retained segments, and date on which the record was last updated. Platform metadata could not be retrieved for 393 uploads (0.67%); the corresponding source addresses were not publicly accessible when metadata retrieval was attempted. These uploads remain represented in `mapping.csv` and `crowd_index.jsonl` because their source identifiers, segment boundaries, locality assignments, and manually curated labels are available; fields derived from `mapping_metadata.csv` are empty for the affected records. Source titles, descriptions, and chapter text are excluded from the public file. The channel identifier is retained to group uploads originating from the same source and assess source concentration; it is not a channel name, handle, or custom URL. The platform upload date should not be interpreted as the recording date. The schema is provided in Table 7.

The `crowd_index.jsonl` file is a derived, segment-level representation of the structured information contained in the two operational CSV tables. It was added to provide a valid JSON representation of fields that are stored as text-encoded or nested lists in `mapping.csv`. Each non-empty line is an independent JSON object representing one retained segment. Nested values are represented using standard JSON objects and arrays, allowing users to read the index without applying Python-specific parsing to the CSV fields. Each object contains the source upload identifier and address, a unique segment identifier, segment start and end

Table 9: Summary distribution of technical properties measured from all 58,467 source videos analysed using `ffprobe` during dataset construction. Video share is calculated relative to all analysed source videos, while duration share is based on their source video durations rather than the retained CROWD segment duration. Only the derived technical property metadata and summary results are released; the source videos were not retained after processing and are not redistributed.

Property	Category	Videos	Video share (%)	Duration share (%)
Exact resolution	1280×720	54,950	93.98	95.19
Exact resolution	960×720	1,056	1.81	0.64
Exact resolution	1280×676	285	0.49	0.63
Exact resolution	1270×720	271	0.46	0.60
Exact resolution	640×360	207	0.35	0.39
Exact resolution	Remaining exact resolutions, each $< 0.33\%$	1,698	2.90	2.56
Resolution bucket	720p	57,408	98.19	98.44
Resolution bucket	480p	633	1.08	0.63
Resolution bucket	360p	276	0.47	0.46
Resolution bucket	2160p	123	0.21	0.41
Resolution bucket	240p	14	0.02	0.02
Resolution bucket	1080p	10	0.02	0.02
Resolution bucket	1440p	2	0.00	0.01
Resolution bucket	144p or lower	1	0.00	0.00
Frame rate	30 fps	27,682	47.35	44.83
Frame rate	29.97 fps	20,323	34.76	37.34
Frame rate	25 fps	4,557	7.79	8.65
Frame rate	24 fps	2,574	4.40	3.89
Frame rate	59.94 fps	1,124	1.92	1.97
Frame rate	60 fps	856	1.46	1.77
Frame rate	Non-standard frame rates, each $< 0.52\%$	1,351	2.31	1.55
Aspect ratio	16:9	55,848	95.52	97.01
Aspect ratio	4:3	1,293	2.21	0.85
Aspect ratio	320:169	287	0.49	0.63
Aspect ratio	Remaining exact aspect ratios, each $< 0.32\%$	1,039	1.78	1.51
Codec	h264	56,545	96.71	96.76
Codec	vp9	1,574	2.69	2.67
Codec	av1	348	0.60	0.57
Compression proxy	moderate bpppf	50,545	86.45	83.58
Compression proxy	low bpppf, strongly compressed	7,303	12.49	15.20
Compression proxy	high bpppf	333	0.57	0.53
Compression proxy	very low bpppf, highly compressed	273	0.47	0.66
Compression proxy	very high bpppf	13	0.02	0.03

Table 10: Results of the automated release validation performed on the final frozen CROWD dataset.

Validation result	Value	Status
Mapping rows checked	8,049	PASS
Unique source uploads referenced	58,467	PASS
Metadata rows checked	58,074	PASS
Uploads without retrievable platform metadata	393	WARN
JSONL records checked	75,147	PASS
JSONL invalid records	0	PASS
JSONL duplicate segment identifiers	0	PASS
JSONL missing segments	0	PASS
JSONL extra segments	0	PASS
JSONL records inconsistent with mapping	0	PASS
Achieved JSONL segment coverage (%)	100.00	PASS
Expected retained segments	75,147	PASS
Detection CSV files found	75,147	PASS
Detection CSV files matched to segments	75,147	PASS
Detection CSV files semantically checked	75,147	PASS
Detection rows semantically checked	All rows	PASS
Missing detection CSV files	0	PASS
Unmatched or unrecognised detection CSV files	0	PASS
Duplicate segment-to-file associations	0	PASS
Achieved detection segment coverage (%)	100.00	PASS
Invalid bounding-box rows	0	PASS
Exact duplicate detection rows	0	PASS
Non-monotonic frame-index rows	0	PASS
Rows with missing track identifiers	0	PASS
Inconsistent track-identifier rows	0	PASS
Validator warning checks	1	WARN
Validator failures	0	PASS

times, the manual time-of-day label, the apparent recording-platform label, the platform upload date where available, a nested locality object, limited structured provenance metadata, and the source mapping row number. Its schema is provided in Table 8.

For each retained segment, the release provides a CSV file named `{video_id}_{start_time}_{fps}.csv`. Each row contains an automatic object detection and its associated segment-local track identifier, confidence score, normalised bounding-box coordinates, object class, and frame index. The `yolo-id` field contains the MS COCO class index produced by YOLO11x. Bounding boxes are represented by normalised centre coordinates, width, and height. The `unique-id` field contains the BoT-SORT track identifier, which is unique only within the corresponding segment and cannot be compared across segments. These files contain automatic detection and tracking pseudo-labels and should not be interpreted as manually validated ground truth. Their schema is provided in Table 11, and the tracking configuration is documented in Table 12.

Technical validation

To ensure internal consistency and release integrity, we implemented an automated release *validator* that operates directly on `mapping.csv`, `mapping_metadata.csv`, `crowd_index.jsonl`, and the per-segment detection CSV files. It checks structural correctness, cross-file consistency, segment coverage, duplicate associations, and row-level detection semantics. It does not constitute a manual assessment of the accuracy of the detector or tracker. The validator was run on the final frozen dataset with full row-level semantic checking enabled and completed without failures. It produced one coverage warning reporting the 393 uploads for which platform metadata could not be retrieved when the corresponding source addresses were found to be unavailable. The final results are summarised in Table 10.

The final run checked all 8,049 locality records in `mapping.csv`, all 58,074 available upload records in `mapping_metadata.csv`, and all 75,147 segment records in `crowd_index.jsonl`. It also found and semantically checked all 75,147 per-segment detection CSV files. Each detection file was matched with its corresponding retained segment, with no missing, unmatched, or duplicate file associations. Both JSONL and detection-file segment coverage were 100%. The 393 unavailable platform-metadata records do not reduce segment coverage because the affected uploads retain their segment definitions and manual annotations in `mapping.csv`.

Downloaded video format and compression descriptors. Because video technical properties can affect perception performance, we characterised all 58,467 downloaded source video files associated with the processing inventory. For each file, we extracted the first video stream metadata using `ffprobe`, including frame width, frame height, average frame rate, codec, pixel format, duration, and file size. All 58,467 files were probed successfully, with no failed probes. Table 9 reports the directly observed distributions by resolution, frame rate, aspect ratio, codec, and compression proxy. Video share is calculated relative to all 58,467 analysed files, while duration share is calculated from their source video durations.

Because CROWD is sourced from YouTube, the original camera recordings prior to YouTube encoding and compression were unavailable. The downloaded files were already compressed platform versions and therefore could not be used as references for calculating reference based perceptual quality metrics. We instead report bits per pixel per frame (bpppf) as an approximate compression proxy, computed as effective bitrate divided by frame width $\times$ frame height $\times$ frame rate. Effective bitrate was calculated as file size $\times 8$ divided by duration, rather than relying on stream level bitrate fields that may be missing or implausible. We grouped this proxy into five broad categories: very low bpppf (< 0.03), low bpppf (0.03 to < 0.06), moderate bpppf (0.06 to < 0.12), high bpppf (0.12 to < 0.25), and very high bpppf (≥ 0.25). This measure should be interpreted as a technical descriptor of compression level rather than as perceptual quality ground truth. Across the 58,467 analysed files, the median effective bitrate was 1.99 Mbps (interquartile range 1.81 to 2.10 Mbps), and the median bpppf was 0.0729 (interquartile range 0.0665 to 0.0773).

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, platform labels, and segment boundary lists, are consistent. It also checks that categorical fields use the permitted values. Within each

Table 11: Schema of per segment YOLO detection and BoT-SORT tracking pseudo-label CSV 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.

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, invalid categorical values, and invalid segment definitions; they confirm structural completeness of the released labels but do not turn the labels into fine grained perception ground truth.

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 uploads, the number of upload records, the number of segment records, and the total annotated duration obtained by summing the adjusted processing interval, $t_{\text{end,processed}} - t_{\text{start}}$, 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.

Validation of the JSON Lines index. The validator reads crowd_index.jsonl line by line and verifies that every non-empty line is valid JSON, that every record is an object with a unique non-empty segment_id, and that the required top-level and nested location fields are present. Each JSONL segment is compared with the corresponding segment derived from mapping.csv, including the upload identifier, start and end times, source mapping row, time-of-day code, and apparent platform code. The final run parsed all 75,147 JSONL records successfully. It found no invalid JSON lines, duplicate segment identifiers, missing required fields, missing or extra segments, or values inconsistent with the corresponding mapping records. The achieved JSONL segment coverage was 100%.

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 reports mapping-table uploads without a corresponding metadata record and metadata records whose upload identifier is absent from the mapping, excluding placeholder identifiers. The final run identified 393 of 58,467 unique mapping-table uploads without a metadata record; these source addresses were no longer publicly accessible when platform metadata retrieval was attempted, and the result is reported as a coverage warning rather than a failure of the retained segment annotations. For the 58,074 available metadata records, the validator also checks that stored per-video segment counts are consistent with the number of unique segments derived from the mapping table after deduplicating segments by their start and end boundaries.

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 final run found 75,147 detection CSV files and

matched every file to one of the 75,147 expected retained segments. It found no missing files, unmatched or unrecognised files, or duplicate segment-to-file associations, yielding 100% detection-file segment coverage.

Semantic validation of detection and tracking rows. For the full release run, the validator opened all 75,147 uniquely matched detection CSV files and checked that all required columns were present and readable. For every detection row, it verified finite and permitted class identifiers, confidence values in $[0, 1]$, positive normalised box dimensions, box extents within the image bounds, and valid positive frame indices. It also checked for exact duplicate rows, decreases in frame index in file order, missing or invalid track identifiers, changes in object class within a track, and reuse of the same track identifier more than once in the same frame. The validator found zero invalid bounding-box rows, zero exact duplicate rows, zero non-monotonic frame-index rows, zero rows with missing track identifiers, and zero inconsistent track-identifier rows.

Table 12: BoT-SORT tracking hyperparameters used to generate segment level track identifiers in the released per segment CSV files.

Hyperparameter	Value	Description
`tracker_type`	`botsort`	Tracker implementation selected by the custom YAML configuration.
`track_high_thresh`	0.70	Confidence threshold for the first association stage during tracking.
`track_low_thresh`	0.30	Confidence threshold for the second association stage during tracking.
`new_track_thresh`	0.70	Minimum confidence required to initialise a new track when no matches are found.
`track_buffer`	$2 \times$ FPS frames	Frame-rate-dependent runtime value corresponding to approximately two seconds. The YAML value of 60 at the 30 fps reference rate is overwritten before segment processing.
`match_thresh`	0.60	Similarity threshold used to match existing tracks with detections.
`fuse_score`	`true`	Combines detection confidence with IoU distance for matching.
`gmc_method`	`sparseOptFlow`	Global motion compensation method used for moving camera footage.
`proximity_thresh`	0.50	IoU threshold for spatial proximity during association.
`appearance_thresh`	0.25	Threshold for appearance based matching.
`with_reid`	`true`	Enables appearance based re identification.
`model`	`auto`	No separate ReID weights were specified; resolution of `auto` followed Ultralytics 8.4.46.

Usage Notes

The dataset distributes the structured `crowd_index.jsonl` index, the source CSV tables, platform-assigned video and channel identifiers, segment start and end times, limited structured provenance metadata, and derived detection and tracking pseudo-labels, but it does not redistribute the underlying video, frames, audio, thumbnails, channel names or handles, or source free text. Users can access referenced uploads through YouTube using the provided source addresses or video identifiers. For analyses that require local video processing, we provide the code used in this work to retrieve and preprocess the referenced uploads in the Code Availability section; use of this code is subject to platform terms and video availability. Because the referenced videos are hosted by YouTube, some uploads and metadata may become unavailable or change over time due to removal, restriction, regional limitations, or account changes.

Software, device, and deterministic execution. The supporting package versions were torchvision 0.23.0, NumPy 1.26.4, SciPy 1.11.4, `lap` 0.5.12, `filterpy` 1.4.5, and PyYAML 6.0.3. Video retrieval and inspection used `pytubefix` 10.10.1, `yt-dlp` 2026.7.4, and FFmpeg and `ffprobe` 4.4.2. PyTorch was built with CUDA 12.8 support, and the inspected environment reported cuDNN 9.10.2. The complete installed package manifest is provided with the reproducibility outputs.

The processing code selects CUDA when available and otherwise falls back to CPU. The generated environment report for the inspected machine records an NVIDIA GeForce RTX 4080 SUPER GPU with 16 GB of memory and NVIDIA driver 560.35.05. This machine record is not presented as the hardware used for every historical processing job because the GPU model was not contemporaneously logged for all earlier runs. The Ultralytics configuration used seed 0 and enabled its deterministic option. However, PyTorch deterministic algorithms and deterministic cuDNN execution were not enabled, and determinism related environment variables, including `CUBLAS_WORKSPACE_CONFIG` and `PYTHONHASHSEED`, were unset. The processing pipeline is therefore not claimed to be bitwise deterministic across hardware or software environments.

Privacy and data protection. CROWD should be treated as a privacy-sensitive research index because the provided source identifiers and segment boundaries may allow users to access videos containing identifiable people, vehicles, residences, and locations. The released files do not contain video frames, audio, thumbnails, source free text, face identities, number-plate transcriptions, persistent person identities, facial embeddings, biometric templates, or walking-gait descriptors. Platform-assigned channel identifiers are retained only for grouping uploads by source, assessing source concentration, detecting duplicate source associations, and auditing provenance; they should not be used to identify or profile uploaders. Downstream users who retrieve the referenced videos create their own processing context and are responsible for complying with platform terms, copyright, privacy, data protection, and research ethics requirements. We recommend data minimisation, restricted access to any locally retrieved video, avoidance of individual identification, and task-specific institutional review where source footage or derived imagery is processed.

Users should be aware of several limitations relevant to reuse. The object detections and tracks are automatic pseudo-labels generated 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). They are not validated annotations and should not be treated as ground truth for object detection or tracking accuracy. These pseudo-labels do not include lane markings, road geometry, or a comprehensive traffic sign taxonomy beyond the COCO categories. Detection and tracking quality may degrade under occlusion, dense traffic, motion blur, low illumination, adverse weather, unusual camera viewpoints, and domain shifts across countries or localities. Track identifiers are not propagated across segment boundaries because tracking is reinitialised per segment. Users requiring ground truth labels should validate or manually annotate the relevant subset for their task.

The coverage of time-of-day is uneven: most retained content is driving during the day (Table 4). 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). CROWD also separates the platform upload date from the recording date: only upload dates are provided, and no independently verified recording dates are released. Users should therefore avoid interpreting Figure 2 as a temporal distribution of driving exposure or using upload dates directly for policy, seasonal, or Covid-19 analyses unless they add independently supported recording date evidence with explicit provenance and uncertainty.

Time-of-day and platform labels. The `time_of_day` field provides one binary contextual label for every retained segment. Its operational definition, intended scope, and limitations are described earlier in the Methods section. The field supports broad filtering, descriptive stratification, and dataset characterisation. Users requiring finer illumination categories or uncertainty-aware labels should independently relabel or audit the relevant subset for their task.

The `vehicle_type` field should likewise be interpreted as an apparent recording platform or a category of a confirmed platform source, not as an independently verified platform identity. The appearance of the platform can be uncertain in forward-facing video sourced from the web because the height of the camera, the mounting geometry, the field of view, the reflections, the shadows, and the source descriptions may be incomplete or ambiguous. Automated driving status is not independently verified in CROWD; automated variants were assigned only when the title, description, or chapter information provided by the

Table 13: Pinned software environment and effective inference settings used to generate the released per-segment detection and tracking pseudo-label files.

Item	Value	Meaning
Python	3.12.13	Python runtime used to execute the detection and tracking pipeline.
Ultralytics	8.4.46	Software package used to load the YOLO11x model and run detection and tracking inference.
PyTorch	2.8.0+cu128	Deep learning framework used to execute the YOLO11x neural network.
OpenCV	4.11.0.86	Library used for video decoding, frame handling, and related image operations.
Image size	640	Nominal inference input size passed to Ultralytics. Video frames were internally resized and padded for model inference while preserving their aspect ratio.
NMS IoU threshold	0.70	Overlapping detections of the same class were suppressed when their intersection over union exceeded this threshold during non maximum suppression.
Detector confidence threshold	0.00	Minimum confidence threshold passed to Ultralytics during YOLO11x inference. Subsequent detection association used the BoT-SORT thresholds reported separately.
Maximum detections	300	Maximum number of object detections retained for any processed frame after non maximum suppression.
Class agnostic NMS	false	Non-maximum suppression was performed separately for each object class, so boxes assigned to different classes did not suppress one another.
Class filter	none	No object classes were excluded during inference; detections from all 80 MS COCO classes were eligible for retention.
Rectangular inference	false	Standard square inference batches were used rather than minimum rectangle batching based on the aspect ratios of the input frames.
Half precision	false	Inference used full 32 bit floating point calculations rather than 16 bit floating point calculations.
DNN inference	false	The OpenCV DNN inference backend was not used; model inference was performed through PyTorch.
Augmentation	false	Each frame was processed once without test time augmentation such as scaling or flipping.
Video stride	1	Every frame in each retained segment was processed, with no systematic frame skipping.

source indicated automated driving or autonomous shuttle operation and should be treated as apparent or asserted by the source unless a downstream user independently verified them.

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). The concentration at the country and locality level is also substantial (Table 3): the ten largest contributing countries account for 54.05% of retained hours, while the top 1%, 5% and 10% of localities account for 66.32%, 84.16% and 89.09% of retained hours, respectively. Studies that aim to make claims across regions should therefore consider continent, country, and locality specific sampling strategies, minimum duration thresholds, weighting, and stratified reporting.

Data Availability

The CROWD dataset is available via 4TU.ResearchData (`https://doi.org/10.4121/06e9bb9a-a064-412b-b0f3-9ac5dd62ea16`). The deposit includes `crowd_index.jsonl`, `mapping.csv`, `mapping_metadata.csv`, and the per-segment detection and tracking CSV files. The deposited index files, manual labels, contextual and limited structured provenance metadata, derived detection and tracking pseudo-labels, and validation outputs are released under a CC BY 4.0 data licence. The source free text and the underlying YouTube videos are not redistributed, are not covered by this licence, and remain governed by platform terms, copyright, privacy, and data protection requirements, and the uploader-selected licence terms. The source code and configuration files are provided separately under the MIT License, as described in the Code Availability section.

Code Availability

The source code to download videos, curate segments, generate `crowd_index.jsonl`, produce the per-segment pseudo-label CSVs, and run the release validator is available at `https://github.com/crowd-dataset/crowd` and is released under the MIT License. The repository contains the JSONL conversion script, `validation_release.py`, the release configuration files (e.g. `config`, `bbox_custom_tracker.yaml`), and an environment specification (`pyproject.toml`) targeting Python 3.12.13. The reproducibility outputs include `reproducibility_report.json`, the complete installed package manifest `pip_freeze.txt`, the model and tracker checksums, the effective inference settings, the tracker configuration, and the recorded CUDA, cuDNN, GPU, and deterministic settings. The validator also regenerates the machine-readable validation report, detailed issue table, manuscript-ready validation table, and checksum manifest. The repository contains the `ffprobe`-based video format and compression summary script used to produce Table 9, and the FFmpeg scene-change screen described in the Technical validation section.

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] European Parliament and Council of the European Union. Regulation (eu) 2016/679 (general data protection regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, 2016. Article 89: Safeguards and derogations relating to processing for scientific or historical research purposes.

[60] 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 Ying Fang, 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 (`https://www.tue.nl/en/research/institutes/eindhoven-artificial-intelligence-systems-institute/digital-twin-lab`), and the TU/e Supercomputing Center (HPC) (`https://supercomputing.tue.nl/`) 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.