MicroVision: An open dataset and benchmark models for detecting vulnerable road users and micromobility vehicles

Safety Science 205 (2027) 107371

Contents lists available at ScienceDirect

Safety Science

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

MicroVision: An open dataset and benchmark models for detecting vulnerable road users and micromobility vehicles☆

Alexander Rasch *, Rahul Rajendra Pai

Department of Mechanical Engineering, Chalmers University of Technology, Hörsalsvägen 7A, 412 58, Gothenburg, Sweden

ARTICLE INFO

Dataset link: https://doi.org/10.71870/eepz-jd52, https://github.com/microlab-chalmers/microvision-code

Keywords:
Micromobility
Open dataset
Traffic safety
Images
Bicycle
E-scooter
Object detection

ABSTRACT

Micromobility is a growing mode of transportation, raising new challenges for traffic safety and planning due to increased interactions in areas where vulnerable road users (VRUs) share the infrastructure with micromobility, including parked micromobility vehicles (MMVs). Approaches to support traffic safety and planning increasingly rely on detecting road users in images—a computer-vision task relying heavily on the quality of the images to train on. However, existing open image datasets for training such models lack focus and diversity in VRUs and MMVs, for instance, by categorizing both pedestrians and MMV riders as "person", or by not including new MMVs like e-scooters. Furthermore, datasets are often captured from a car perspective and lack data from areas where only VRUs travel (sidewalks, cycle paths). To help close this gap, we introduce the MicroVision dataset: an open image dataset and annotations for training and evaluating models for detecting the most common VRUs (pedestrians, cyclists, e-scooterists) and stationary MMVs (bicycles, e-scooters), from a VRU perspective. The dataset, recorded in Gothenburg (Sweden), consists of more than 8000 anonymized, full-HD images with more than 30,000 carefully annotated VRUs and MMVs, captured over an entire year and part of almost 2000 unique interaction scenes. Along with the dataset, we provide first benchmark object-detection models based on state-of-the-art architectures, which achieved a mean average precision of up to 0.726 on an unseen test set. The dataset and model can support traffic safety to distinguish between different VRUs and MMVs, or help monitoring systems identify the use of micromobility.

1. Introduction

Micromobility refers to a mode of transport typically represented by compact and lightweight vehicles operating at lower speeds compared to more traditional modes of transport like passenger cars (S.A.E. International, 2019). Micromobility has become a rising trend globally, particularly for personal transportation (Olabi et al., 2023). Its rise has been explained by its compactness and cost-effectiveness — often making it a popular last-mile option — but also by improvements to health and sustainability, as it is a more environmentally friendly mode of transportation (McQueen et al., 2021; Olabi et al., 2023).

While the use of micromobility has increased over the last few years, so have interactions with other road users, as well as crashes (Yannis et al., 2024). As most users of micromobility vehicles (MMVs) fall within the category of vulnerable road users (VRUs), they are less protected than users of motorized vehicles such as cars or trucks. Consequently, VRUs are more likely to be injured in collisions, particularly when they do not wear helmets or collide with heavier motorized vehicles. As a response to critical interactions with motorized vehicles, VRUs may also become psychologically discouraged from continuing to travel, which may have negative long-term consequences for society. Therefore, improving the safety of micromobility is paramount to further promote and sustain its use.

Following the safe systems approach (Khan and Das, 2024), initiatives to promote safe micromobility focus on road-user behavior, infrastructure, speed, and vehicle design. Initiatives focusing on behavior, for instance, study interactions between road users to understand what makes these interactions critical and to model behavior so that other initiatives can predict it and become more effective. For larger motorized vehicles, both active and passive safety systems have been developed to prevent crashes with VRUs or mitigate their consequences (Silla et al., 2017). However, for MMVs, such approaches have been scarce, partly due to tighter budget constraints and the limited space available for sensors and actuators.

A common challenge across all approaches is the need to accurately detect and distinguish micromobility vehicles and users in data. Such distinction is important because previous research has shown that e-scooterists, cyclists, and pedestrians have unique characteristics, such

☆ This article is part of a Special issue entitled: 'Micromobility Safety' published in Safety Science.

* Corresponding author.
E-mail addresses: alexander.rasch@chalmers.se (A. Rasch), rahul.pai@chalmers.se (R.R. Pai).

https://doi.org/10.1016/j.ssci.2026.107371
Received 20 March 2026; Received in revised form 31 May 2026; Accepted 9 July 2026
Available online 7 September 2026

as differences in travel speed, route choice, and braking distance during obstacle avoidance (Dozza et al., 2022, 2023; Gilroy et al., 2022; Kazemzadeh et al., 2023; Pai and Dozza, 2025). For instance, an e-scooterist may appear from afar like a pedestrian, but travel at speeds comparable to a cyclist (della Mura et al., 2022). Detecting VRUs has become particularly important for image data captured by cameras, which are central to modern safety systems due to their relatively low cost, small size, and ability to capture the environment in a manner similar to human vision.

Image-based object-detection models for VRUs have been typically combined with distance or depth-estimation models that can localize detected objects in 3D space and thereby enable the precise tracking of road users and the identification of safety-critical events (Zhang et al., 2025). To enhance localization accuracy, camera data are often fused with LIDAR or radar data, which complement the camera images' richness in contextual details with highly accurate distance information. Alternatively, other approaches have used the detected object coordinates in the image directly. García-Venegas et al. (2021), for instance, defined five specific zones in the image and detected risky situations depending on which of these zones the cyclist was detected in, combined with the detected heading angle. Bharti (2023) and Fang and Wu (2024) trained machine learning using ground-truth LIDAR data to predict the VRU's distance and angle from the camera based on their image coordinates. Tracking the VRU's position then allowed identifying safety-critical events using time-to-collision (TTC). All of these approaches share the need for high-accuracy object detection to avoid detecting false-positive events. However, developing object-detection models with high accuracy requires large amounts of diverse training data.

2. Related work

Previous work has provided multiple open traffic datasets that have proved useful to the research community developing models for detecting road users and vehicles (Alibeigi et al., 2023; Geiger et al., 2013; Sun et al., 2020). However, these datasets have primarily been captured from a car perspective, thereby lacking data from cycle paths or sidewalks where only VRUs can travel. Furthermore, such datasets often miss finer distinctions between different types of micromobility, such as bicycles and e-scooters, or omit e-scooters entirely. E-scooters are known to perform differently compared to bicycles (Distefano et al., 2024; Dozza et al., 2022); therefore, distinguishing them from bicycles is important. In addition, detecting not only VRUs but also stationary MMVs has become increasingly important, as parked vehicles may pose obstacles to other road users.

Existing work on VRU detection has predominantly focused on pedestrians and cyclists. Most studies have developed annotated image datasets — some openly available — and trained deep-learning-based object-detection models. Model architectures have largely been based on *convolutional neural networks* (CNNs), which can be broadly categorized into two main classes: (1) region-based (two-stage) detectors, such as R-CNN, Fast R-CNN, and Faster R-CNN, which first generate region proposals that are subsequently classified; and (2) single-shot (one-stage) detectors, such as YOLO and SSD, which directly predict object locations and classes (García-Venegas et al., 2021). While two-stage detectors generally achieve higher accuracy at the cost of slower inference, single-stage detectors are often preferred for applications requiring faster inference, potentially at the expense of accuracy. With the advent of *transformers* in computer vision, transformer-based detectors such as DETR, RT-DETR, and RF-DETR have gained attention for their end-to-end formulation and strong performance in complex scenes, albeit with higher computational demands.

Previous research on cyclist detection has predominantly relied on CNN-based detection models. García-Venegas et al. (2021), for example, developed an image dataset and detection model for cyclists and their orientations, concluding that region-based detectors yielded more accurate predictions, while single-shot detectors offered faster but less accurate results that could be sufficient when combined with object tracking. Xiaofei Li et al. (2016) presented a public cyclist dataset from China containing more than 20,000 annotated instances and reported consistent detection performance across several then state-of-the-art architectures, including Fast R-CNN.

Due to the relatively recent appearance of e-scooters compared to bicycles on public roads, research on detecting e-scooters in images has been limited. Apurv et al. (2021) were among the first to present a public image dataset including e-scooters in the United States, consisting of approximately 10,000 images from 83 interaction scenes. Along with the dataset, they introduced a MobileNetV2-based benchmark model to distinguish e-scooter riders from pedestrians. Gilroy et al. (2022) further improved rider classification by training a model on web-sourced images of partially occluded e-scooter riders. Chen et al. (2024) published a dataset of approximately 2000 images collected in Charlottesville, USA, with over 11,000 annotated objects, and presented benchmark detection models using one-stage YOLO architectures. Sabri et al. (2024) introduced another public dataset of stationary MMVs, based on web-sourced videos featuring bicycles, e-scooters, and skateboards, and proposed a YOLOX-based detector augmented with temporal features, achieving improved detection accuracy.

Few studies have addressed the direct detection of e-scooterists, defined as the combined entity of rider and vehicle, and most have focused instead on image-level classification to distinguish them from cyclists (Apurv et al., 2021; Gilroy et al., 2022). Additionally, many datasets have been collected over relatively short time spans, potentially missing seasonal variations in environmental conditions and road-user appearance. Most available datasets including e-scooterists have also been recorded in North America (Apurv et al., 2021; Chen et al., 2024; Sabri et al., 2024). To support better generalization of detection models, additional datasets from other regions, including Europe, are needed (Alibeigi et al., 2023).

In this study, we address the lack of data for micromobility safety by introducing *MicroVision*, an open object-detection dataset accompanied by initial benchmark models. The dataset comprises over 8000 high-resolution anonymized images with more than 30,000 annotated instances of VRUs and MMVs. Unlike existing car-centric datasets, the images are captured from the ego perspective of micromobility users, covering VRU infrastructure often absent from other datasets and spanning a full annual cycle in Gothenburg, Sweden. Furthermore, we introduce a state-aware classification scheme that explicitly distinguishes between active riders (e.g., e-scooterists) and stationary vehicles (e.g., parked e-scooters), a distinction critical for downstream safety tasks such as trajectory prediction and risk assessment. Finally, we provide benchmark models based on state-of-the-art object-detection architectures to support future research on micromobility detection, particularly for traffic-safety applications.

3. Method

3.1. Data collection

The images were derived from video footage systematically collected within the city of Gothenburg, Sweden. Data acquisition spanned a full annual cycle (from July 2024 to June 2025) and multiple times of day, thereby capturing a wide range of seasonal and lighting conditions to ensure substantial environmental and temporal diversity. Footage was captured from a micromobility user's perspective by recording the surroundings while (1) riding an e-scooter, (2) riding a bicycle, and (3) walking. The ego-perspective recording strategy was chosen to address a gap in existing datasets, which are predominantly captured from a fixed or car-centric perspective. The viewpoints align with the growing deployment of onboard cameras on shared micromobility fleets (Pai and Dozza, 2025).

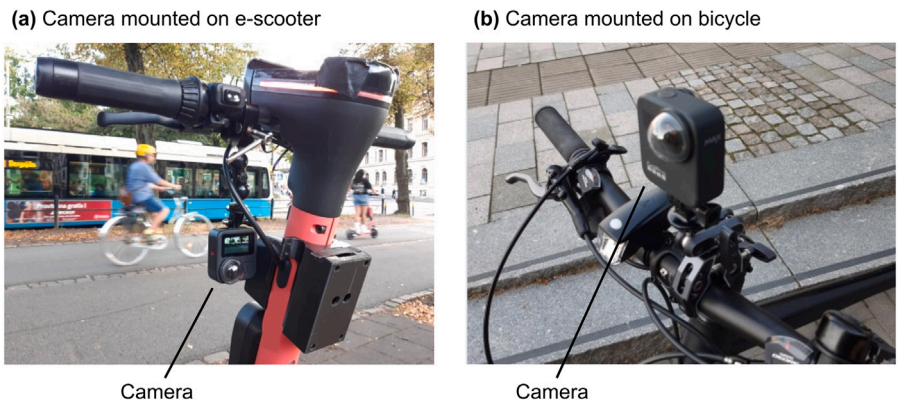

Fig. 1. Placement of the GoPro Max camera on the e-scooter (a) and bicycle (b) used for data collection.

Data acquisition covered a diverse range of urban infrastructure, including dedicated cycle paths, sidewalks, shared pedestrian-cyclist zones, and signalized intersections, thereby capturing the variety of environments in which VRUs typically travel. A GoPro Max camera was used for all recordings. For vehicle-based recordings, the camera was mounted on the handlebar at heights varying between 1.0 and 1.5 m above the ground (see Fig. 1); while walking, the camera was handheld. Videos were recorded at a resolution of 1920 × 1080 pixels at 60 frames per second. The high frame rate ensured sharp captures of moving road users at close proximity. To minimize lens distortion and provide images suitable for standard object-detection architectures without intensive post-calibration, the camera was operated in a *linear* lens mode.

The systematic recording of public spaces raises ethical concerns regarding the incidental collection of personal data without explicit individual consent. Requesting consent from each road user is impractical and may affect the realism of captured scenes. To address this, the data collector displayed a clearly visible sign during all recording sessions informing nearby road users about the ongoing data collection and providing contact information for inquiries or data-removal requests. Prior to public release, all data underwent anonymization to obscure identifiable personal information and comply with local data-protection regulations. Specifically, faces and vehicle license plates were blurred using the Precision Blur service provided by brighter AI, following prior work (Alibeigi et al., 2023). The collection protocol and data provision were approved by the Swedish Ethical Review Authority (reference number 2024-00329-01).

3.2. Frame extraction

From the recorded videos, relevant frames were extracted for annotation. The extraction aimed to maximize environmental diversity while maintaining a balanced distribution of object classes. Only a limited number of frames depicting the same object from different angles were selected. This process was facilitated by pre-annotating the full videos with the current best model (YOLO11-based, fine-tuned from public COCO-pretrained weights). To maintain consistent object identities across frames, we employed the BoT-SORT tracker (Aharon et al., 2022). Tracking objects across frames and selecting frames sufficiently spaced in time ensured coverage of different viewpoints. Each candidate frame was then manually verified to confirm visual diversity and to exclude near-duplicate scenes.

3.3. Annotations

Selected frames were annotated with 2D rectangular bounding boxes for common VRUs and MMVs. Five classes were defined: (1) *pedestrian*, (2) *cyclist*, (3) *e-scooterist*, (4) stationary *bicycle*, and (5) stationary *e-scooter*. Consistent with the National Highway Traffic Safety Administration definition, a pedestrian was defined as a person who is walking or sitting, provided they were not on a vehicle (National Center for Statistics and Analysis, 2021). For micromobility, we adopted a state-aware labeling strategy: a cyclist or e-scooterist was defined as the combination of the vehicle and at least one person traveling with it, including cases where the rider was stationary (e.g., waiting at a traffic light) or mounting the vehicle. Conversely, a person pushing a vehicle was labeled as two separate objects: a pedestrian and a vehicle. Stationary vehicles were annotated regardless of their orientation (upright or fallen).

Tricycles (e.g., bicycles with two front wheels) were included in the bicycle category in accordance with Swedish vehicle classification standards. Mopeds and large cargo cycles used for deliveries were excluded due to their rare appearance to avoid dataset imbalance. Smaller attachments (e.g., child seats or backpacks) were included within bounding boxes, while larger pushed or pulled objects (e.g., suitcases or child buggies) were excluded. Broken vehicles (e.g., missing wheels) were included in MMV categories. Following the COCO dataset conventions (Lin et al., 2014), occluded objects were annotated with boxes covering the visible extent; in cases of severe occlusion (e.g., dense bicycle racks), only the closest and most distinguishable object was annotated. Full documentation of the annotation format, directory structure, class encoding, and dataset splits, together with sample code for loading and training, can be reached through the project page of the dataset.[1]

To maximize efficiency and consistency, a model-assisted semi-automated annotation workflow was employed (Fig. 2). First, an object-detection model (YOLO7; Wang et al. (2022)) was fine-tuned on a small initial set of open-source web images (Fang and Wu, 2024). This model was used to pre-annotate the first batch of images. Three human annotators then reviewed and corrected labels using Label Studio v1.13.1 (Tkachenko et al., 2025). The corrected data were used to fine-tune a new model based on the same YOLO7 architecture for pre-annotating subsequent batches, which was eventually updated with a YOLO11 architecture for improved performance (Jocher and Qiu, 2024). This iterative predict-correct-retrain cycle was repeated until the entire dataset was processed, progressively reducing manual effort.

A final model-guided quality check was conducted to identify and correct potential human annotation errors. The best-performing model from the iterative loop was fine-tuned on the full dataset (90%/10% train/validation split) and used to predict labels for all images. Discrepancies between predictions and ground truth were manually reviewed, focusing on false negatives and false positives, and corrected as needed. This validation was repeated with varying splits until no labeling issues were observed.

To mitigate model bias toward object-dense scenes and reduce false positives in empty environments, a small number of background images

[1] https://microlab-chalmers.github.io/microvision-dataset/

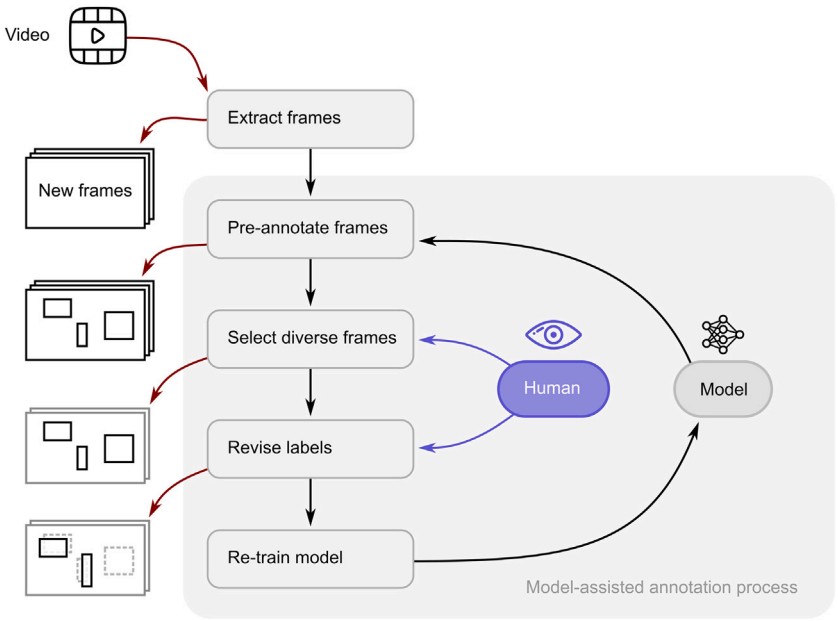

Fig. 2. Overview of the data processing pipeline, from raw videos to revised labels used for model training.

(approximately 5% of the dataset) containing none of the defined classes were added. These images were identified by the model as having no detections and verified manually.

3.4. Inter-annotator agreement

To assess reliability and consistency, an inter-annotator agreement analysis was conducted on a randomly selected subset of 258 images independently labeled by the three annotators. Pairwise agreement metrics were computed for all annotator pairs and averaged. For each pair, annotations were matched using the Hungarian algorithm to maximize total Intersection over Union (IoU) for bounding boxes of the same class (Kuhn, 1955). Matches were considered valid only if IoU exceeded 0.5, consistent with common object-detection benchmarks such as Pascal VOC (Everingham et al., 2010). Based on valid matches, *box agreement* was computed as the average IoU to quantify spatial precision. *Class agreement* was assessed using Cohen's Kappa (Cohen, 1960), accounting for chance agreement. Metrics were stratified by object class and object size to ensure robustness across road-user types and distances.

3.5. Benchmark object-detection models

To provide initial performance benchmarks, three state-of-the-art object-detection architectures representing distinct paradigms were trained and evaluated: (1) one-stage YOLO version 11 (YOLO11; (Jocher and Qiu, 2024)), (2) two-stage Faster R-CNN (Ren et al., 2015), and (3) transformer-based RF-DETR (Robinson et al., 2025). Publicly available pretrained weights were fine-tuned for all models. To focus on architectural comparison rather than hyperparameter optimization, a uniform training setup was used. All models were trained for 100 epochs with an effective batch size of 32 and an input resolution of 1280 pixels (for RF-DETR, the nearest compatible resolution of 1232 pixels was used). YOLO11 (largest available model variant "X") was trained using the Ultralytics package (v8.3.103). Faster R-CNN (largest variant "X101-FPN") was trained using the Detectron2 package (v0.6; Wu et al. (2019)). RF-DETR (largest variant "large") was trained using the rfdetr package (v1.3.0) provided by Roboflow. Training was performed using up to four parallel NVIDIA A100 GPUs (80 GB VRAM each) on the Alvis compute cluster provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS).

To avoid data leakage — that is, to prevent the same physical objects from appearing in both training and evaluation splits — the dataset was split by *scenes* rather than by individual frames. Consecutive frames from the same scene may depict the same objects from slightly different angles, and thereby inflate reported performance by allowing models to memorize specific object appearances. Scenes were defined as temporal sequences containing the same objects, identified using YOLO11 tracking outputs; transitions between scenes were detected when the last tracked object disappeared for more than 10 s before the next appeared. Scenes were stratified into training (80%), validation (10%), and test (10%) sets.

Model performance was evaluated on the held-out test set using mean average precision (mAP), both per class and averaged across classes (Beitzel et al., 2009). Following common practice, mAP was computed as the average over IoU thresholds from 0.5 to 0.95 in increments of 0.05 (mAP@0.5:0.95). In addition, performance was analyzed by object size following the COCO protocol (Lin et al., 2014). Given the higher image resolution (1920 pixels wide) compared to COCO, size thresholds were scaled proportionally: *small* (area $< 96^2$ px^2), *medium* (96^2 $px^2 \leq$ area $< 288^2$ px^2), and *large* (area $\geq 288^2$ px^2). For the CNN-based detectors YOLO11 and Faster R-CNN, we chose a default non-maximum suppression (NMS) threshold of 0.5.

4. Results

4.1. The MicroVision dataset

The MicroVision dataset contains 8706 images, of which 594 are background images without any annotated objects. The remaining 8113 images are annotated with a total of 34,866 objects using 2D bounding boxes. These include 17,032 pedestrians, 4772 cyclists, 5091 e-scooterists, 4289 bicycles, and 3682 e-scooters (see Fig. 3 for some examples). The images are organized into 1984 unique scenes. Fig. 4 shows the distributions of bounding-box widths and heights for all classes. Overall, the dataset consists of 67% small, 25% medium, and 8% large objects. The comparatively large number of pedestrians, particularly small pedestrians, is a consequence of recording in dense urban environments, which increases background pedestrian frequency. While MMVs (bicycles and e-scooters) are predominantly represented by smaller and lower bounding boxes, VRU bounding boxes tend to be taller, as they combine both vehicle and rider.

Fig. 3. Example images and annotations from the MicroVision dataset for different object classes (columns) and object sizes (rows).

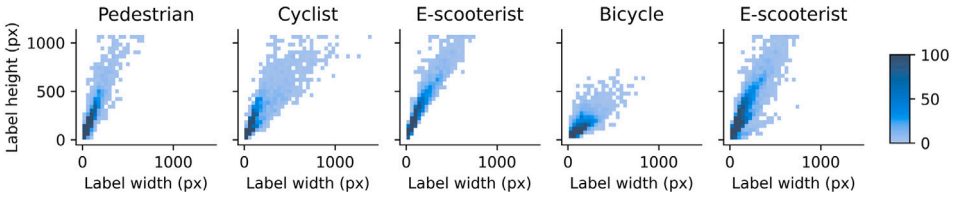

Fig. 4. Bounding-box width and height distributions for the different object classes.

Table 1
Inter-annotator agreement for different object classes and object sizes (S = small, M = medium, L = large). Class agreement is reported using Cohen's Kappa, and box agreement using mean Intersection over Union (IoU).

Class	Class agreement (Cohen's Kappa)				Box agreement (mean IoU)			
	S	M	L	All	S	M	L	All
Pedestrian	0.794	0.943	0.939	0.811	0.857	0.914	0.967	0.866
Bicycle	0.676	0.837	1.000	0.761	0.872	0.891	0.955	0.886
Cyclist	0.873	0.989	0.967	0.922	0.897	0.942	0.974	0.922
E-scooter	0.670	0.837	0.897	0.728	0.847	0.904	0.968	0.873
E-scooterist	0.877	0.987	0.978	0.938	0.911	0.953	0.957	0.937
All classes	0.783	0.919	0.963	0.824	0.866	0.922	0.964	0.887

The inter-annotator agreement analysis yielded an overall Cohen's Kappa of 0.824 and an average Intersection over Union (IoU) of 0.887. Agreement correlated positively with object size, increasing from small objects (Kappa = 0.783; IoU = 0.866) to large objects (Kappa = 0.963; IoU = 0.964). Across classes, active road users (cyclists and e-scooterists) exhibited higher annotation consistency than stationary vehicles, as summarized in Table 1.

4.2. Object-detection benchmark

Among the evaluated models, RF-DETR achieved the best overall performance on the held-out test set, with an overall mAP of 0.726, followed by YOLO11 (0.687) and Faster R-CNN (0.580), as reported in Table 2. While RF-DETR excelled in overall performance for each object size and class, and in particular for medium and large objects, YOLO11 outperformed the other models on small pedestrians and cyclists, small and medium-sized e-scooters, as well as small e-scooterists, almost closing the performance gap to RF-DETR for small objects overall. Faster R-CNN performed consistently worse across all object classes and object sizes. Image anonymization resulted in negligible performance differences across all models (YOLO11: 0.690 original vs. 0.687 anonymized; Faster R-CNN: 0.579 original vs. 0.580 anonymized; RF-DETR: 0.731 original vs. 0.726 anonymized).

Fig. 5 illustrates ground-truth annotations and predictions from the two best-performing models (YOLO11 and RF-DETR) on representative test images using an IoU threshold of 0.6 and a confidence threshold of 0.5. Fig. 6 highlights common failure cases. RF-DETR demonstrated superior performance for partially visible objects, such as occluded objects or objects near image boundaries. In some cases, both models incorrectly included adjacent objects (e.g., suitcases pulled by pedestrians) within bounding boxes; RF-DETR occasionally misclassified such instances. YOLO11, in contrast, correctly detected pedestrians pushing bicycles but sometimes produced redundant cyclist detections, indicating potential limitations related to NMS. Additional metrics and confusion matrices are reported in Table A.1 and Fig. A.1, respectively.

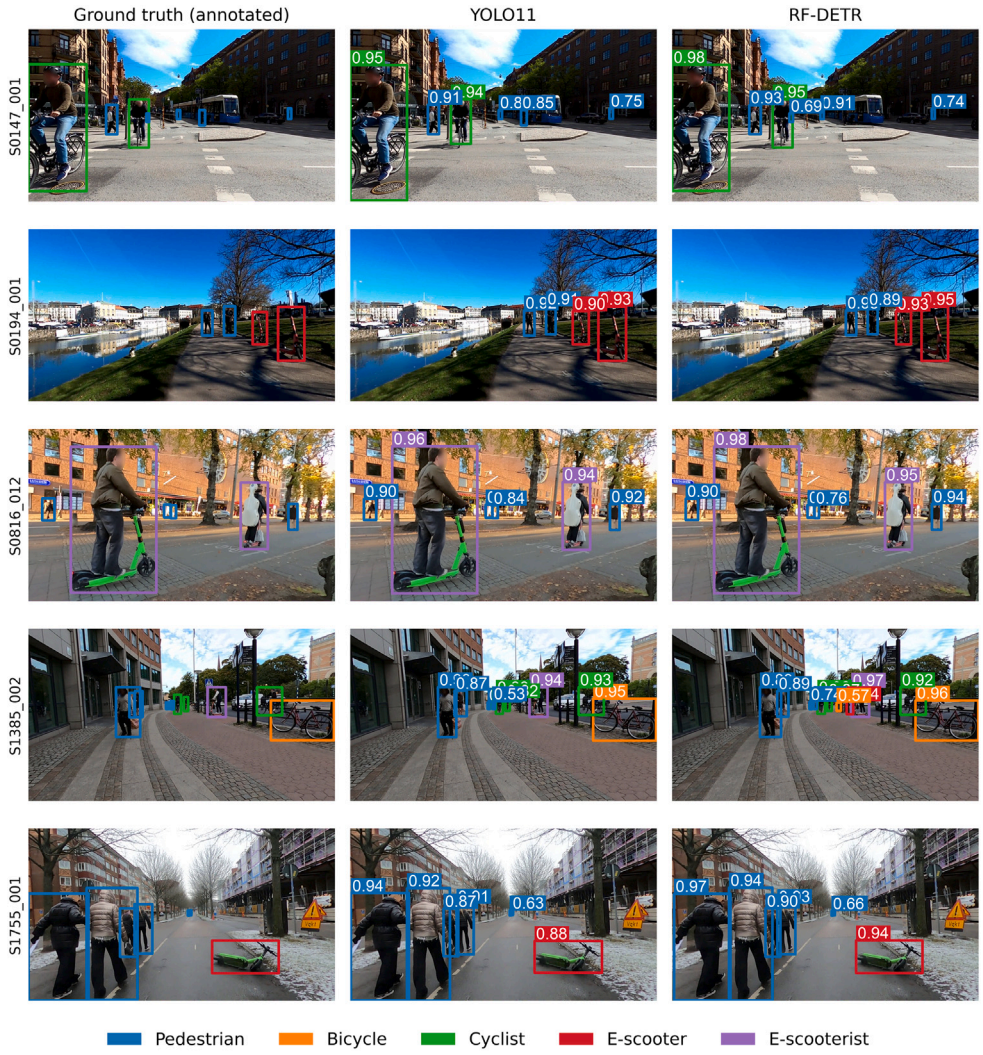

Fig. 5. Successful example predictions on images from unseen test-set scenes for YOLO11 and RF-DETR, shown alongside ground-truth annotations. Numbers indicate predicted confidence scores (threshold = 0.5).

5. Discussion

5.1. Dataset and benchmark models

The MicroVision dataset provides a specialized resource for understanding complex interactions in urban micromobility environments. While its scale of 8706 images is smaller than many established car-centric datasets (Alibeigi et al., 2023; Geiger et al., 2013; Sun et al., 2020), its scientific value lies in its qualitative uniqueness and technical specificity. Most existing open traffic datasets are recorded from the perspective of passenger cars and capture scenes primarily from the center of motorized traffic lanes. In contrast, the MicroVision dataset is captured directly from the perspective of micromobility users and pedestrians. This viewpoint shift is critical for developing safety systems intended for MMVs themselves, as it captures the specific visual context of sidewalks, cycle paths, and other shared urban spaces that are largely absent or underrepresented in car-centric data.

Furthermore, the dataset introduces a state-aware classification strategy that explicitly distinguishes between active riders (e.g., cyclists or e-scooterists) and stationary vehicles (e.g., bicycles or e-scooters). This distinction is essential for traffic safety and trajectory prediction, as a vehicle with a rider represents a dynamic entity with immediate kinetic potential and intent, whereas a parked vehicle constitutes a static obstacle. By providing high-resolution data collected over a full annual cycle, MicroVision ensures that these classifications are robust to environmental and seasonal variations typical of a Northern European urban setting. The large number of interaction scenes (nearly 2000) further contributes to diversity in road-user appearance, viewing angles, and surrounding environments.

The initial benchmarking of state-of-the-art object-detection models demonstrated promising performance across architectures on unseen scenes. The transformer-based RF-DETR model outperformed CNN-based architectures (YOLO11 and Faster R-CNN) in overall performance, achieving an mAP of 0.726 and excelling particularly at detecting partially visible and occluded objects. This performance gap suggests that attention mechanisms may offer superior scene comprehension in dense urban environments. However, this improved accuracy comes at increased computational cost: the evaluated RF-DETR variant contains approximately 135 million parameters and requires longer inference times, making it less suitable for latency-critical real-time applications compared to more lightweight models such as YOLO11 (57 million parameters). All evaluated models struggled with reliably detecting stationary MMVs, largely due to dense clustering in parking zones where vehicles occlude one another. Such cases pose challenges not only for automated detection but also for consistent human annotation, as reflected in the inter-annotator agreement analysis.

Model performance further degraded for small and distant objects, such as far-away pedestrians, which is a well-known limitation in

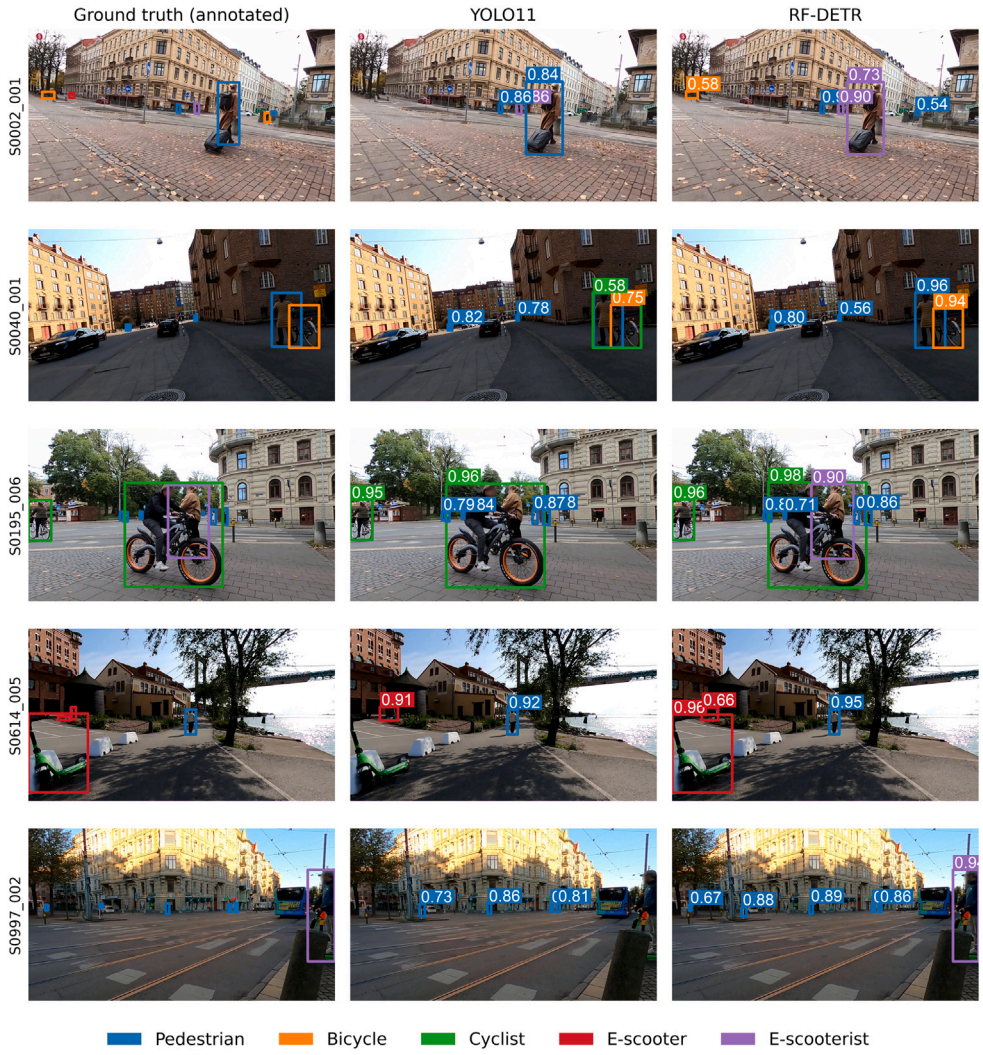

Fig. 6. Representative examples of prediction errors from YOLO11 and RF-DETR, shown alongside ground-truth annotations.

object detection. From a traffic-safety perspective, immediate collision risks are typically lower for such distant objects; however, applications requiring long-range perception may benefit from complementary techniques. Object tracking can mitigate missed detections over time, and Slicing Aided Hyper Inference (SAHI) can improve detection of small objects by performing inference on image subregions (Akyon et al., 2022). Additional challenging cases included fallen or deformed (e.g., folded) MMVs, which were comparatively rare in the dataset and therefore more difficult for models to learn reliably.

5.2. Implications and applications

Within the transportation domain, the MicroVision dataset enables a wide range of applications. Traffic-safety research and development can leverage the dataset and benchmark models to build advanced driver or rider assistance systems for both motorized vehicles and MMVs, aiming to prevent collisions with VRUs and MMVs. Detection outputs can serve as an input to downstream tasks such as trajectory prediction that account for distinct behavioral patterns, for example, the greater unpredictability often associated with e-scooters compared to bicycles (Distefano et al., 2024); however, such analyses require additional processing steps beyond frame-level detection, including object tracking across frames to establish temporal correspondence, as well as depth or distance estimation to recover metric spatial information. Surrogate safety measures such as TTC cannot be derived from static 2D detections alone.

Beyond real-time systems, the benchmark models can be combined with depth estimation to support retrospective traffic-safety research by enabling automated identification of road-user interactions in large-scale naturalistic video datasets (Beck et al., 2019; Dozza and Werneke, 2014; Pai and Dozza, 2025; Schleinitz et al., 2017; Guo et al., 2026; Fang and Wu, 2024). Such datasets are typically annotated manually, a process that is both time-consuming and error-prone. The MicroVision dataset and models may facilitate behavioral research by enabling efficient discovery of specific road-user interactions, appearances, and conflict scenarios.

Although primarily designed for safety-related research, the dataset may also benefit traffic management and urban planning. Automated detection and counting of VRUs can support infrastructure assessment, demand modeling, and policy evaluation (Peixoto et al., 2020; Vacalebri et al., 2025). However, the relatively low mounting height of the camera, corresponding to a road-user perspective, may limit the direct applicability of trained models to imagery captured from elevated viewpoints, such as building-mounted cameras.

5.3. Limitations and future work

While the dataset captures a broad range of objects and environmental conditions, it is geographically limited to Gothenburg, Sweden. As

Table 2
Comparison of YOLO11, Faster R-CNN, and RF-DETR on the test set using COCO mAP@[0.5:0.95]. S, M, and L denote small, medium, and large objects, respectively. Bold values indicate the best performance per class and object size.

Class	Model	S	M	L	All
Pedestrian	YOLO11	**0.442**	0.645	0.769	0.597
	Faster R-CNN	0.279	0.559	0.726	0.499
	RF-DETR	0.438	**0.673**	**0.869**	**0.629**
Bicycle	YOLO11	0.110	0.373	0.724	0.518
	Faster R-CNN	0.131	0.280	0.633	0.431
	RF-DETR	**0.233**	**0.411**	**0.818**	**0.597**
Cyclist	YOLO11	**0.353**	0.700	0.883	0.766
	Faster R-CNN	0.093	0.570	0.818	0.669
	RF-DETR	0.332	**0.725**	**0.932**	**0.813**
E-scooter	YOLO11	**0.232**	**0.609**	0.837	0.692
	Faster R-CNN	0.052	0.408	0.694	0.510
	RF-DETR	0.226	0.583	**0.879**	**0.702**
E-scooterist	YOLO11	**0.574**	0.787	0.927	0.865
	Faster R-CNN	0.280	0.671	0.880	0.789
	RF-DETR	0.499	**0.816**	**0.950**	**0.889**
All classes	YOLO11	0.342	0.623	0.828	0.687
	Faster R-CNN	0.167	0.497	0.750	0.580
	RF-DETR	**0.346**	**0.641**	**0.890**	**0.726**

a result, the generalization of trained models to regions with different infrastructure designs, traffic regulations, or vehicle types remains an open question. Moreover, most images were collected during daytime and under favorable weather conditions, reflecting periods of higher VRU activity and practical constraints on data collection. Safely operating a micromobility vehicle during adverse weather conditions while simultaneously collecting data posed safety risks, and camera image quality degrades substantially under low-light and precipitation conditions, reducing annotation reliability. Future work should investigate whether additional data are required to improve generalization to nighttime or adverse weather conditions. Nonetheless, the dataset provides a strong foundation for future model-assisted annotation workflows, enabling efficient extension to new domains.

Regarding scope, the annotation effort focused on the most prevalent micromobility forms currently observed in Sweden. Less common VRUs and MMVs, such as low-speed mopeds or monowheels, were not included, nor were fine-grained distinctions within categories (e.g., bicycles with trailers). Future dataset expansions could address these gaps by extending temporal and spatial coverage and including motorized vehicle classes. Although general-purpose detectors trained on car-centric datasets can identify common motorized vehicles, models specifically trained from a micromobility perspective may offer improved performance for the viewpoints and interaction scenarios characteristic of VRU infrastructure. Additionally, extending annotations beyond 2D bounding boxes to include instance segmentation could enable more precise pixel-level scene understanding and support tasks such as drivable-space estimation.

6. Conclusion

We present an open image dataset with annotations and a first object-detection benchmark to advance the detection of vulnerable road users and micromobility vehicles. Comprising over 8000 high-resolution images from nearly 2000 unique interaction scenes and more than 30,000 annotations, the dataset addresses a critical gap by capturing the visual contexts of sidewalks and cycle paths and by employing a state-aware annotation strategy that distinguishes active riders from stationary vehicles.

Benchmarking state-of-the-art object-detection architectures shows that transformer-based models such as RF-DETR achieve superior accuracy in complex and occluded scenes, while efficient single-stage detectors like YOLO11 offer a favorable trade-off for real-time applications. Together, the dataset and benchmark models provide a foundation for next-generation traffic systems that can improve the safety and comfort of micromobility users, both through real-time assistance systems and by enabling partial automation of video-based behavioral research. By releasing these resources openly, we aim to support the development of models that generalize across the diverse and evolving landscape of urban transportation with a focus on micromobility.

CRediT authorship contribution statement

Alexander Rasch: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rahul Rajendra Pai:** Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Shiyi Qiu, Mahin Garg, and Anton Broman (Chalmers University of Technology) for their assistance with data processing and annotation, and Marco Dozza (Chalmers University of Technology) for valuable discussions and funding acquisition. The authors also thank the Chalmers Data Office for support with the practical and administrative aspects of dataset preparation and publication.

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

This work was carried out within the *MicroVision* project, funded by Vinnova (Sweden's innovation agency), the Swedish Energy Agency, and Formas (the Swedish Research Council for Sustainable Development) through the DriveSweden program (reference number 2023-01047).

Appendix. Additional evaluation metrics

See Fig. A.1.
See Table A.1.

Data availability

The MicroVision dataset and benchmark model weights are publicly available via an access request at https://doi.org/10.71870/eepz-jd52, hosted on the Data Organization and Information System (DORIS) by the Swedish National Data Service (SND). Code to reproduce the analyses and data-processing pipeline are available at https://github.com/microlab-chalmers/microvision-code.

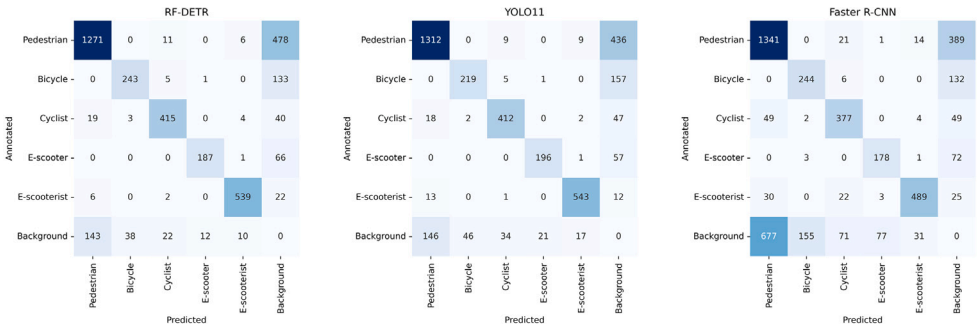

Fig. A.1. Confusion matrices for the three models, evaluated on the test set, shown for an example IoU threshold of 0.6 and a confidence threshold of 0.5.

Table A.1
Comparison of the models on the test set for the metrics accuracy (A), precision (P), recall (R), and F1-score (F1), using an IoU threshold of 0.6 and a confidence threshold of 0.5 as an example.

Class	Model	Acc.	Prec.	Rec.	F1
Pedestrian	RF-DETR	0.657	**0.883**	0.720	0.793
	YOLO11	**0.675**	0.881	0.743	**0.806**
	Faster R-CNN	0.532	0.639	**0.759**	0.694
Bicycle	RF-DETR	**0.574**	**0.856**	0.636	**0.730**
	YOLO11	0.509	0.820	0.573	0.675
	Faster R-CNN	0.450	0.604	**0.639**	0.621
Cyclist	RF-DETR	**0.797**	**0.912**	**0.863**	**0.887**
	YOLO11	0.777	0.894	0.857	0.875
	Faster R-CNN	0.627	0.759	0.784	0.771
E-scooter	RF-DETR	0.700	**0.935**	0.736	0.824
	YOLO11	**0.710**	0.899	**0.772**	**0.831**
	Faster R-CNN	0.531	0.687	0.701	0.694
E-scooterist	RF-DETR	**0.914**	**0.963**	0.947	**0.955**
	YOLO11	0.908	0.949	**0.954**	0.952
	Faster R-CNN	0.790	0.907	0.859	0.883
All Classes	RF-DETR	**0.711**	**0.904**	0.769	**0.831**
	YOLO11	0.710	0.892	**0.777**	0.830
	Faster R-CNN	0.569	0.693	0.762	0.725

References

Aharon, N., Orfaig, R., Bobrovsky, B.-Z., 2022. Bot-SORT: Robust associations multi-pedestrian tracking. arXiv:2206.14651URL https://arxiv.org/abs/2206.14651.

Akyon, F.C., Onur Altinuc, S., Temizel, A., 2022. Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. In: 2022 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 966–970. http://dx.doi.org/10.1109/ICIP46576.2022.9897990, arXiv:2202.06934.

Alibeigi, M., Ljungbergh, W., Tonderski, A., et al., 2023. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. http://dx.doi.org/10.48550/arXiv.2305.02008.

Apurv, K., Tian, R., Sherony, R., 2021. Detection of E-scooter riders in naturalistic scenes. arXiv:2111.14060, URL https://arxiv.org/abs/2111.14060.

Beck, B., Chong, D., Olivier, J., et al., 2019. How much space do drivers provide when passing cyclists? Understanding the impact of motor vehicle and infrastructure characteristics on passing distance. Accid. Anal. Prev. 128, 253–260. http://dx.doi.org/10.1016/j.aap.2019.03.007.

Beitzel, S.M., Jensen, E.C., Frieder, O., 2009. MAP. In: LIU, L., ÖZSU, M.T. (Eds.), Encyclopedia of Database Systems. Springer US, Boston, MA, pp. 1691–1692. http://dx.doi.org/10.1007/978-0-387-39940-9_492.

Bharti, K., 2023. Estimating Road-User Position from a Camera: a Machine Learning Approach to Enable Safety Applications. URL https://odr.chalmers.se/items/c5c843b1-773e-4c9d-a8a7-b291ca6614fd.

Chen, D., Hosseini, A., Smith, A., et al., 2024. Performance Evaluation of Real-Time Object Detection for Electric Scooters. arXiv:2405.03039.

Cohen, J., 1960. A coefficient of agreement for nominal scales.. Educ. Psychol. Meas. 20, 37–46. http://dx.doi.org/10.1177/001316446002000104.

Distefano, N., Leonardi, S., Kieć, M., et al., 2024. Comparison of E-scooter and bike users' behavior in mixed traffic. Transp. Res. Rec. http://dx.doi.org/10.1177/03611981241263339.

Dozza, M., Li, T., Billstein, L., et al., 2023. How do different micro-mobility vehicles affect longitudinal control? Results from a field experiment. J. Saf. Res. 84, 24–32. http://dx.doi.org/10.1016/j.jsr.2022.10.005.

Dozza, M., Violin, A., Rasch, A., 2022. A data-driven framework for the safe integration of micro-mobility into the transport system: Comparing bicycles and e-scooters in field trials. J. Saf. Res. 81, 67–77. http://dx.doi.org/10.1016/j.jsr.2022.01.007.

Dozza, M., Werneke, J., 2014. Introducing naturalistic cycling data: What factors influence bicyclists' safety in the real world? Transp. Res. Part F: Traffic Psychol. Behav. 24, 83–91. http://dx.doi.org/10.1016/J.TRF.2014.04.001.

Everingham, M., Van Gool, L., Williams, C.K.I., et al., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88 (2), 303–338. http://dx.doi.org/10.1007/s11263-009-0275-4.

Fang, L., Wu, Y., 2024. Threat assessment from naturalistic video-data: How to detect, classify, and estimate the position of multiple road users from cameras. URL https://odr.chalmers.se/items/1f59fb35-7b6a-49b6-958e-d0fe540917ee.

García-Venegas, M., Mercado-Ravell, D.A., Pinedo-Sánchez, L.A., et al., 2021. On the safety of vulnerable road users by cyclist detection and tracking. Mach. Vis. Appl. 32 (5), 1–17. http://dx.doi.org/10.1007/s00138-021-01231-4.

Geiger, A., Lenz, P., Stiller, C., et al., 2013. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research. Int. J. Robot. Res. (October), 1–6. http://dx.doi.org/10.1177/0278364913491297.

Gilroy, S., Mullins, D., Jones, E., et al., 2022. E-Scooter Rider detection and classification in dense urban environments. Results Eng. 16 (October), 100677. http://dx.doi.org/10.1016/j.rineng.2022.100677, arXiv:2205.10184.

Guo, J., Cheng, Y., Wang, M., et al., 2026. Multiple object detection and tracking in panoramic videos for cycling safety analysis. IET Intell. Transp. Syst. 20 (1), e70228. http://dx.doi.org/10.1049/itr2.70228.

Jocher, G., Qiu, J., 2024. Ultralytics YOLO11. URL https://github.com/ultralytics/ultralytics.

Kazemzadeh, K., Haghani, M., Sprei, F., 2023. Electric scooter safety: An integrative review of evidence from transport and medical research domains. Sustain. Cities Soc. 89 (November 2022), 104313. http://dx.doi.org/10.1016/j.scs.2022.104313.

Khan, M.N., Das, S., 2024. Advancing traffic safety through the safe system approach: A systematic review. Accid. Anal. Prev. 199 (February), 107518. http://dx.doi.org/10.1016/j.aap.2024.107518.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Nav. Res. Logist. Q. 2 (1–2), 83–97. http://dx.doi.org/10.1002/nav.3800020109.

Lin, T.Y., Maire, M., Belongie, S., et al., 2014. Microsoft COCO: Common objects in context. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 8693 LNCS (PART 5), 740–755, arXiv:1405.0312.

McQueen, M., Abou-Zeid, G., MacArthur, J., et al., 2021. Transportation transformation: Is micromobility making a macro impact on sustainability? J. Plan. Lit. 36 (1), 46–61. http://dx.doi.org/10.1177/0885412220972696.

della Mura, M., Failla, S., Gori, N., et al., 2022. E-scooter presence in urban areas: Are consistent rules, paying attention and smooth infrastructure enough for safety? Sustain. (Switzerland) 14 (21), http://dx.doi.org/10.3390/su142114303.

National Center for Statistics and Analysis, 2021. Pedestrians: 2019 data (Traffic Safety Facts. Report No. DOT HS 813 079). Tech. Rep. May, National Highway Traffic Safety Administration, URL https://crashstats.nhtsa.dot.gov/Api/Public/Publication/813079.

Olabi, A.G., Wilberforce, T., Obaideen, K., Sayed, E.T., Shehata, N., Alami, A.H., Abdelkareem, M.A., 2023. Micromobility: Progress, benefits, challenges, policy and regulations, energy sources and storage, and its role in achieving sustainable development goals. Int. J. Thermofluids 17 (January), 100292. http://dx.doi.org/10.1016/j.ijft.2023.100292.

Pai, R.R., Dozza, M., 2025. Understanding factors influencing e-scooterist crash risk: A naturalistic study of rental e-scooters in an urban area. Accid. Anal. Prev. 209 (June 2024), 107839. http://dx.doi.org/10.1016/j.aap.2024.107839.

Peixoto, E., Moutinho, J., José, R., 2020. A low-cost video-based solution for city-wide bicycle counting in starter cities. In: Santos, H., Pereira, G.V., Budde, M., Lopes, S.F., Nikolic, P. (Eds.), Science and Technologies for Smart Cities. Springer International Publishing, Cham, pp. 139–150. http://dx.doi.org/10.1007/978-3-030-51005-3_14.

Ren, S., He, K., Girshick, R., et al., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. vol. 28, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

Robinson, I., Robicheaux, P., Popov, M., et al., 2025. RF-DETR: Neural architecture search for real-time detection transformers. arXiv:2511.09554.

Sabri, K., Djilali, C., Bilodeau, G.-A., et al., 2024. Detection of micromobility vehicles in urban traffic videos. Proc. Conf. Robot. Vis. (Mmv), http://dx.doi.org/10.21428/d82e957c.abc3243f, arXiv:2402.18503.

S.A.E. International, 2019. Taxonomy and Classification of Powered Micromobility Vehicles, SAE Standard J3194_201911. http://dx.doi.org/10.4271/J3194_201911.

Schleinitz, K., Petzoldt, T., Franke-Bartholdt, L., et al., 2017. The German Naturalistic Cycling Study – Comparing cycling speed of riders of different e-bikes and conventional bicycles. Saf. Sci. 92, 290–297. http://dx.doi.org/10.1016/J.SSCI.2015.07.027.

Silla, A., Leden, L., Rämä, P., et al., 2017. Can cyclist safety be improved with intelligent transport systems? Accid. Anal. Prev. 105, 134–145. http://dx.doi.org/10.1016/j.aap.2016.05.003.

Sun, P., Kretzschmar, H., Dotiwalla, X., et al., 2020. Scalability in perception for autonomous driving: Waymo open dataset. Proc. the IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2443–2451. http://dx.doi.org/10.1109/CVPR42600.2020.00252, arXiv:1912.04838.

Tkachenko, M., Malyuk, M., Holmanyuk, A., et al., 2025. Label Studio: Data labeling software. URL https://github.com/HumanSignal/label-studio.

Vacalebri, F., Moll, S., López, G., et al., 2025. AI-enhanced road safety: Real-time information on cycle traffic on rural roads. In: Juan, A.A., Faulin, J., Lopez-Lopez, D. (Eds.), Decision Sciences. Springer Nature Switzerland, Cham, pp. 281–288. http://dx.doi.org/10.1007/978-3-031-78241-1_26.

Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M., 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696.

Wu, Y., Kirillov, A., Massa, F., et al., 2019. Detectron2. URL https://github.com/facebookresearch/detectron2.

Xiaofei Li, Flohr, F., Yue Yang, et al., 2016. A new benchmark for vision-based cyclist detection. In: 2016 IEEE Intelligent Vehicles Symposium. IV, 2016-Augus, (Iv), IEEE, pp. 1028–1033. http://dx.doi.org/10.1109/IVS.2016.7535515.

Yannis, G., Petraki, V., Crist, P., 2024. Safer Micromobility: Technical Background Report The International Transport Forum. Tech. Rep. March, URL https://www.itf-oecd.org/sites/default/files/safer-micrombility-technical-report.pdf.

Zhang, Z., Wei, C., Wu, G., Barth, M.J., 2025. Vulnerable road user detection for roadside-assisted safety protection: A comprehensive survey. Appl. Sci. 15 (7), http://dx.doi.org/10.3390/app15073797.