Fine-Grained Classification of Vulnerable Road Users: A Robust Track-Level Approach for Urban Surveillance

Mohamed Jaouad Lahgazi
COSYS - PICSL-L laboratory
Gustave Eiffel University
Marne-la-Vallée, France
mohamed-jaouad.lahgazi@univ-eiffel.fr

Jean-Philippe Tarel
COSYS - PICSL-L laboratory
Gustave Eiffel University
Marne-la-Vallée, France
jean-philippe.tarel@univ-eiffel.fr

***Abstract*—Automated monitoring of Vulnerable Road Users (VRUs) using fixed infrastructure cameras is essential for traffic safety analytics in Intelligent Transport Systems. General person detection is well-studied, yet pre-trained models often fail to distinguish between the main VRU categories such as pedestrians, cyclists, motorcyclists, and personal mobility device (PMD) users. These categories require different safety interventions due to their distinct speed profiles, infrastructure requirements, and injury risks. We present a track-level classification pipeline that combines real-time person detection, multi-object tracking, and a Swin Transformer classifier fine-tuned on a custom four-class VRU dataset. A voting scheme, based on confidence gating, aggregates frame-level predictions into a single per-track label, making the system robust to temporary occlusion and detection noise. We also introduce an evaluation protocol in which the pipeline is run once on the evaluation video, tracks are labeled by a human annotator, and accuracy is measured by direct track identifier, eliminating the spatial re-matching artifacts that affect conventional pipeline evaluation. The pipeline is evaluated on 2,965 annotated tracks drawn from three urban surveillance recordings, achieving an overall accuracy of 99.70% across all four VRU classes. The pipeline, pretrained weights, and evaluation scripts are released publicly to support reproducible research.**

***Index Terms*—vulnerable road users, multi-object tracking, Swin Transformer, fine-grained classification, infrastructure cameras**

I. Introduction

Road traffic injuries remain a critical global public health challenge. According to the 2023 WHO Global Status Report on Road Safety, 1.19 million people died in traffic accidents in 2021, with vulnerable road users accounting for more than half of all fatalities: 23% were pedestrians, 21% were powered two and three-wheeler riders, 6% were cyclists, and 3% were users of micromobility devices such as electric scooters [1]. Notably, the number of cyclist deaths increased by nearly 20% between 2010 and 2021, and overall safety gains for VRUs has been substantially slower than for vehicle occupants. These numbers demonstrate the urgent need for monitoring and analytical tools designed with light and non-motorized road users' safety in mind.

Supported by the ATUVVU project, funded by the French Road Safety Delegation (*Délégation à la Sécurité Routière*) under the 2022 annual call for projects.

Fixed infrastructure cameras at intersections, road segments, and mobility hubs offer an affordable and non-intrusive solution for continuous VRU monitoring. Computer vision systems built on such cameras enable several critical safety functions: adaptive signal control, traffic counting, dangerous behavior detection, and post-hoc accident investigation [2]. Precise identification of the type of road user present in each scene is fundamental for all such downstream tasks, as the risk profiles and the required system responses for a pedestrian, cyclist, motorcyclist, or PMD user differ significantly during interaction with a vehicle. However, standard person detectors, which collapse these distinct categories into a single `person` class, do not provide this critical granularity.

While pedestrian detection has matured [3], [4], the fine-grained multi-class problem remains underexplored. Specialized benchmarks such as the Tsinghua-Daimler dataset [5] demonstrate that cyclists' detection performance is substantially lower than pedestrians', particularly at long distances and under occlusion. Although recent surveys highlight the evolution of roadside sensing [6], [7], the majority of the literature still lacks a unified classification of pedestrians, cyclists, motorcyclists, and personal mobility devices (PMD) users from infrastructure cameras. The PMD class has only recently attracted attention due to the increasing use of electric scooters in urban areas. Prior studies on the subject have focused more on binary detection than multi-class classification [8], [9]. To the best of our knowledge, no prior work addresses the joint four-class VRU problem (pedestrians, cyclists, motorcyclists, and PMD users) within a single unified pipeline operating on infrastructure camera footage.

The temporal dimension of video monitoring offers a significant opportunity to improve system reliability. A moving VRU is a sequence of bounding boxes created by a tracker over dozens to hundreds of frames rather than a single image. Partial occlusions, motion blur, poor lighting, and the small apparent size of distant subjects all cause temporary misclassifications, making individual frame-level classifications

inherently noisy. Therefore, combining frame-level evidence into a single, consistent per-track label is necessary for reliable monitoring, yet no aggregation methodology, specific to VRUs, has been proposed or evaluated in the literature. In this paper, we tackle both classification and aggregation problems. We fine-tune a Swin Transformer Tiny classifier on a custom four-class VRU image dataset, and incorporate it into a full pipeline. After processing video clips from fixed infrastructure cameras, the pipeline uses a confidence-gated scheme, counting only the high-confidence classifications, to give each tracked VRU a class label and confidence score. Swin Transformer Tiny [10] was selected based on a systematic comparison of nine architectures trained under identical conditions, where it achieved the highest accuracy while offering a favorable trade-off between parameter count and inference speed suitable for near-real-time deployment. A full account of this comparison is deferred to a companion benchmark paper. The following are our main contributions:

- A **fine-grained VRU classifier** based on Swin Transformer Tiny [10], fine-tuned on a custom four-class dataset of 82,200 crops from urban surveillance footage, with pretrained weights released publicly.
- A **unified four-class VRU pipeline**, the first to jointly address pedestrians, cyclists, motorcyclists, and PMD riders from infrastructure cameras, integrating YOLO11 [11] detection, BoT-SORT [12] tracking, and a confidence-gated voting scheme for stable per-track label assignment.
- A **clean evaluation protocol** based on direct track identifier lookup that eliminates spatial re-matching artifacts, together with a full public release of pipeline code, annotation tool, and evaluation scripts.[1]

The remainder of this paper is organized as follows. Section II reviews related work on VRU detection, multi-object tracking, and track-level classification. Section III describes the pipeline architecture. Section IV presents the evaluation protocol and validation dataset construction methodology. Section V reports quantitative results on a validation set of annotated urban surveillance clips from three distinct camera locations. Section VI discusses limitations and future work.

II. Related work

A. VRU detection from infrastructure cameras

Intelligent transport systems research has a long history of using fixed cameras to automatically analyze road users. In the survey [2], authors have shown that the majority of practical deployments are restricted to coarse vehicle versus non-vehicle distinctions, and that road-user detection and classification from monocular cameras are central open problems. The study [4] argued that in order to support meaningful safety interventions, pedestrian protection systems must go beyond presence detection to characterize the type, posture, and intent of each road user. This argument is still relevant to the more general VRU classification problem today.

Over the last ten years, the pedestrian detection sub-problem has developed significantly. With the introduction of CityPersons [3], a large-scale pedestrian dataset based on Cityscapes imagery, it has been shown that detectors trained on domain-matched data considerably outperform those trained on generic benchmarks. This result encourages us to use a specially designed VRU dataset instead of reusing COCO-pretrained weights for the four-class problem.

Cyclists have received significantly less attention than pedestrians. Authors in [5] presented the Tsinghua-Daimler Cyclist Benchmark and demonstrated that even the most advanced detectors at the time deteriorated significantly with increasing difficulty, proving that cyclist detection is still an unresolved issue that is separate from pedestrian detection. Further, in [13], LiDAR scans augmented with synthetic depth images have been used to detect cyclists. The method achieved competitive results, but they needed active sensor hardware that isn't found in typical infrastructure deployments. The surveillance literature largely ignores motorcyclists and PMD users as target classes, a gap that our work directly fills.

B. Electric scooter and PMD detection

The rapid growth of shared micromobility services has motivated the initial attempts at automated PMD detection and classification. The main challenge is that general-purpose person detectors consistently fail to distinguish e-scooter riders from pedestrians at the feature level. Therefore, [8] built a dedicated e-scooter dataset and trained a detection pipeline for dense urban environments. Although their research showed that a purpose-trained classifier can reliably separate riders from non-riders, the detection was performed from a vehicle-mounted camera rather than fixed infrastructure, the classification was binary, and it was not integrated into a multi-class VRU framework. The safety-critical motivation for differentiating this class was highlighted in [9], who also suggested a YOLO-based detection system for e-scooter riders to support injury emergency services. Using video-based trajectory extraction in a controlled campus experiment, [14] examined PMD movement dynamics and found that PMD users had different kinematic profiles from both pedestrians. However, none of these works integrate the PMD class alongside pedestrian, cyclist, and motorcyclist classes in a shared classification framework, nor do they address the track-level label assignment problem.

C. Multi-object tracking

The tracking-by-detection paradigm, which independently detects objects in each frame and then associates detections over time, has taken the lead in multi-object tracking for surveillance applications. This paradigm was developed with SORT by the authors in [15], who combined a Kalman filter for motion prediction with the Hungarian algorithm for assignment. They showed that a simple, real-time implementation could compete with much more complex offline techniques on common benchmarks. In order to reduce identity switches under occlusion, the study [16] extended SORT with a deep appearance descriptor at the cost of an additional re-identification network. More recently, [12] proposed BoT-SORT, which combines camera motion compensation, a refined Kalman

[1] https://github.com/MJLahgazi/vru-pipeline

filter state vector encoding aspect ratio directly, and optional ReID features.

BoT-SORT achieved state-of-the-art performance on MOTChallenge (80.5 MOTA, 65.0 HOTA on MOT17) and integrates seamlessly with the Ultralytics framework, making it well-suited for infrastructure deployments subject to camera shake. Critically, tracking-by-detection systems produce spatial trajectories but no semantic class labels, motivating the confidence-gated voting scheme proposed in this work.

D. Classification backbones and track-level aggregation

Recently, vision transformer architectures are becoming increasingly popular in a wide majority of contemporary image classification backbones after the success of ViT [17]. Among these, the Swin Transformer [10] introduced a hierarchical architecture with shifted-window self-attention. This allows the construction of multi-scale feature maps absent in plain ViT and linear computational complexity with respect to image size. Swin Tiny achieves 81.3% top-1 accuracy on ImageNet-1K with 28M parameters, making it a feasible option for applications where inference must run in near-real-time alongside a detection and tracking pipeline. In addition, the aggregating frame-level data into track-level semantic label has been addressed in limited scope. For instance, one related but fundamentally different problem is the intention recognition of cyclists and pedestrians from pose sequences over time, which has been tackled in [18]. The study has demonstrated that the prediction quality improved when temporal context was used instead of a tracked individual. This work encourages the use of track-level inference instead of frame-level inference. However, the label assignment problem for classification is not addressed. Temporal pooling and majority voting are common methods in the action recognition community for combining clip-level predictions [19]. Confidence-aware aggregation has been explored in adjacent vision tasks, such as video person re-identification [20] and tracklet-level multi-view object classification [21], yet these approaches target fundamentally different domains with distinct noise characteristics. In infrastructure VRU monitoring, crops are frequently degraded by motion blur, partial occlusion, or insufficient resolution due to the subject's distance from the camera, with degradation patterns that vary across VRU classes. Our confidence-gated scheme addresses this specifically: only predictions exceeding a confidence threshold contribute to the final vote, ensuring the track label is determined by the frames where the classifier is most certain about the VRU class.

III. Methodology

As illustrated in Fig. 1, our pipeline processes raw infrastructure videos through three sequential stages: detection, tracking, and classification. A final voting module produces a single stable class label per track.

A. Person detection

Person detection is performed using YOLO11m [11], [22], restricted to the *person* class (COCO class index 0) at a confidence threshold of 0.4 and an IoU threshold of 0.5. Frames are fed via an OpenCV [23] capture loop, allowing the pipeline to process proprietary camera formats (including DAV) not natively supported by the YOLO streaming interface.

B. Multi-object tracking

Detected bounding boxes are passed to BoT-SORT [12], which assigns each detection a persistent integer track identifier stable for the lifetime of the trajectory. BoT-SORT is invoked through the Ultralytics interface with default hyperparameters.

C. Training dataset and classifier

The VRU classifier is trained on a custom dataset of cropped bounding box images spanning four classes: pedestrian, cyclist, motorcyclist, and PMD rider. The source footage was collected from 14 distinct urban locations across 20 different camera views, totaling 2083 hours of recording, from which we annotated 82,200 crops with over 20,000 samples per class. Representative samples are shown in Fig. 2. The dataset is not yet publicly released due to GDPR constraints on the source footage, a privacy-anonymized version is under preparation and will be made publicly available alongside a forthcoming benchmark publication. Pretrained classifier weights are released in the interim to support reproducibility.

The backbone is Swin Transformer Tiny [10], selected following a systematic comparison of nine architectures trained under identical conditions on the same dataset, in which it achieved the highest test accuracy while offering a favorable trade-off between parameter count and inference speed. It uses a hierarchical architecture with shifted window self-attention to achieve linear computational complexity with respect to image size, and was preferred over plain ViT [17] on grounds of efficiency and over ResNet-family models on grounds of its superior feature localization on small, context-rich crops. The model accepts 224×224 RGB inputs normalized to ImageNet statistics. The input pipeline resizes each 160×160 crop to 224×224 before inference.

Training uses the AdamW optimizer with an initial learning rate of 1×10^{-4}, weight decay of 0.05, and a cosine annealing schedule over 100 epochs with a 5-epoch linear warm-up. Early stopping is applied with a patience of 15 epochs. The batch size is 64. Data augmentation includes random horizontal flipping, random rotation up to 10°, and color jitter (brightness, contrast, and saturation each ±0.2). Cross-entropy loss is used without label smoothing. The model achieving the highest validation accuracy is retained. All training is performed on a single NVIDIA RTX A5000 GPU (24GB VRAM).

D. Crop extraction

For each detected bounding box (x_1, y_1, x_2, y_2) with width w and height h, a padded crop is extracted by expanding the box by $0.2w$ laterally, $0.3h$ upward, and $0.1h$ downward to capture contextual cues such as wheels and handlebars. The expanded region is symmetrically padded to a square, zero-padded at image boundaries if necessary, and resized to 160×160.

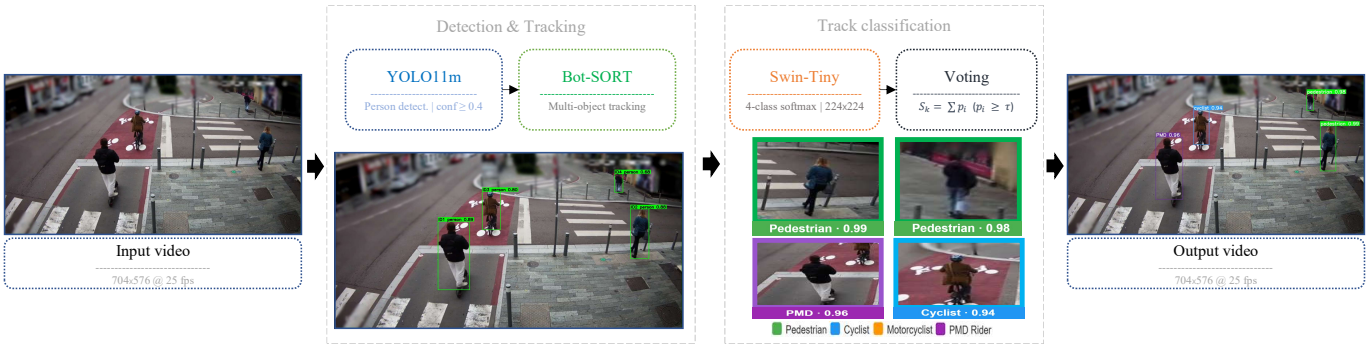


Fig. 1: Overview of the proposed VRU classification pipeline. Frames are processed by YOLO11 for person detection, BoT-SORT for multi-object tracking, and Swin Tiny for per-crop classification. A confidence-weighted voting module aggregates frame-level predictions into a single per-track label.



(a) Pedestrian (b) Cyclist (c) Motorcyclist (d) PMD rider

Fig. 2: Samples from the four-class VRU training dataset. Images are 160 × 160 pixel crops extracted from fixed infrastructure camera footage.

E. Confidence gated voting

Raw frame-level classifier outputs are noisy due to occlusion, motion blur, and the small apparent size of distant road users. Rather than assigning the majority-vote class as the track label, we aggregate predictions using a confidence-gated scheme. Let $\{(c_i, p_i)\}_{i=1}^{N}$ denote the sequence of frame-level predictions for a given track, where $c_i \in \mathcal{C}$ is the predicted class and $p_i \in [0, 1]$ is the softmax confidence. Only predictions with $p_i \geq \tau$ (threshold $\tau = 0.4$) are included in the vote. For each class $k \in \mathcal{C}$, the aggregated score is:

$$S_k = \sum_{\{i\,:\,c_i=k,\ p_i \geq \tau\}} p_i \tag{1}$$

The track label is $\hat{c} = \arg\max_{k \in \mathcal{C}} S_k$ and the reported confidence is $S_{\hat{c}} / \sum_k S_k$. Tracks for which fewer than five frames pass the confidence threshold are marked as *unknown* and excluded from evaluation. This scheme ensures that the final track label is determined by the frames in which the classifier is most certain.

IV. Evaluation protocol and validation dataset

A. Validation dataset construction

The validation set is drawn from three urban surveillance recordings captured at three distinct fixed-camera locations, covering a variety of intersection geometries, traffic densities, and VRU class distributions. Each recording is at 704×576 pixels and 25 fps. The pipeline is run *once* on each video to produce a set of candidate tracks. A quality filter discards tracks shorter than 20 frames and tracks whose median bounding box height is below 60 pixels, removing distant objects whose crops are insufficiently resolved for reliable classification. The remaining tracks from all three recordings are pooled to form the evaluation corpus.

B. Ground-truth annotation

Ground-truth labels are assigned via a an annotation tool we built for this purpose, it has a graphic user interface that displays a grid of keyframe crops for each track and prompts the annotator to assign one of the four VRU classes or to mark the track as *skip*. Skipped tracks, typically arising from severe occlusion or ambiguous appearance, are excluded from all evaluation metrics. Annotations are written to a separate JSON file that records the track identifier and the human-assigned label. Crucially, the annotation file references the same track identifiers produced by the single pipeline run, so ground-truth lookup requires no spatial matching at evaluation time. The annotation tool supports resumable sessions so that large video sequences can be annotated across multiple sittings without risk of duplication.

C. Evaluation metrics

Let $\mathcal{T}_{\text{eval}}$ denote the set of annotated, non-skipped tracks across all three recordings. For each $t \in \mathcal{T}_{\text{eval}}$, the pipeline prediction $\hat{c}_t$ is compared directly to the annotator label c_t by track identifier. No spatial IoU matching or temporal tolerance is required. We report:

- **overall accuracy** $= |\{t : \hat{c}_t = c_t\}| / |\mathcal{T}_{\text{eval}}|$
- per-class **precision**, **recall**, and **F1 score**
- **confusion matrix** over the four VRU classes

It is important to note that this protocol measures *classifier accuracy on quality-filtered tracks* and does not quantify end-to-end detection recall: VRUs not detected by YOLO11, tracks lost by BoT-SORT, or tracks filtered by the quality criteria are not counted as errors here. This is a deliberate scope decision, the evaluation isolates the classification stage, which is the novel contribution of the pipeline, from the detection and tracking stages, whose performance is characterised by existing benchmarks.

V. Experimental results

A. Implementation details

The pipeline runs on a workstation equipped with a single NVIDIA GPU (RTX 5000 ada generation). Detection uses

TABLE I: Per-video and global classification results

Video	Evaluated	Correct	Wrong	Accuracy
Video 1	380	379	1	99.74%
Video 2	297	295	2	99.33%
Video 3	2,288	2,282	6	99.74%
Global	**2,965**	**2,956**	**9**	**99.70%**

TABLE II: Per-class classification results aggregated over three videos ($N = 2{,}965$ evaluated tracks)

Class	Count	Prec.	Rec.	F1
Pedestrian	2,604	99.85%	99.81%	99.83%
Cyclist	258	98.84%	98.84%	98.84%
Motorcyclist	34	100.00%	100.00%	100.00%
PMD	69	97.14%	98.55%	97.84%
Overall	**2,965**		**99.70%**	

YOLO11m pretrained on COCO. The BoT-SORT tracker is used with default Ultralytics configuration. The Swin Tiny classifier was trained from scratch on the four-class VRU dataset described in Section III. All code is implemented in Python using PyTorch [24], Ultralytics [25], and OpenCV [23].

B. Overall classification accuracy

Table I summarizes per-video and global results across the three evaluation recordings. In total, 2,986 tracks were produced by the pipeline across all three videos, 21 were skipped by the annotator due to ambiguous appearance, leaving 2,965 tracks for evaluation. The pipeline achieves an overall accuracy of **99.70%**, with only nine misclassifications across all tracks and all locations.

Table II reports per-class precision, recall, and F1 score aggregated over all three recordings. The pedestrian class accounts for 87.8% of evaluated tracks (2,604 out of 2,965), reflecting the realistic class distribution of urban surveillance footage. The motorcyclist class achieves perfect precision, recall, and F1 across all videos. The cyclist and PMD classes, which are the most visually similar to pedestrians at low resolution, remain above 97% F1.

C. Error analysis

The nine misclassifications observed across all three recordings are consistent with the natural semantic boundaries of the four-class problem. Five pedestrian tracks were misclassified, of which three were assigned to the cyclist class and two to the PMD class. The five correspond to road users whose appearance is transitional in nature: a pedestrian accompanied by a bicycle, for instance, produces handlebars and wheel rims in the bounding box crop that are characteristic of cyclist crops. Three cyclist tracks were misclassified as pedestrians, which is the canonical hard case in the VRU literature: a dismounted cyclist pushing their bicycle presents a pedestrian-like silhouette and gait, and the distinguishing cue is the partially visible bicycle frame rather than a distinctive pose. One PMD track was misclassified as a pedestrian, a plausible confusion at low image resolution when a stationary rider is observed without the vehicle being visible below the crop boundary.

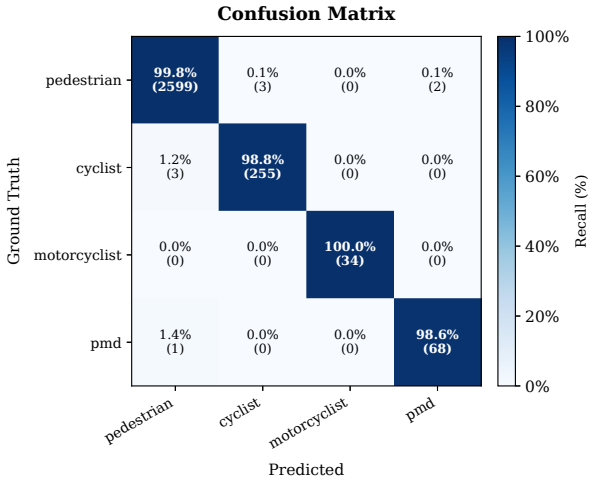


Fig. 3: Normalised confusion matrix aggregated over all three evaluation recordings ($N = 2{,}965$ tracks). Rows denote ground-truth class, columns denote predicted class.

No misclassifications were observed for the motorcyclist class across any of the three recording locations. No confusion was observed between cyclists and motorcyclists, or between PMD riders and motorcyclists, classes that differ substantially in apparent vehicle size and posture. The confusion matrix in Fig. 3 confirms that all nine errors fall on the two natural ambiguity boundaries in the class hierarchy (pedestrian versus cyclist, and pedestrian versus PMD) and that no systematic confusion exists between any other class pair.

D. Effect of the quality filter

Before applying the quality filter (minimum 20 frames, median bounding box height ≥ 60 px), the pipeline produced a larger set of candidate tracks from each evaluation video. The filter removes short, spurious detections and crops of distant road users that are too small for reliable classification, reducing annotation burden while ensuring that the evaluation set consists exclusively of tracks for which correct classification is feasible. The thresholds were chosen to correspond approximately to road users within 15–20 m of the camera on 704×576 footage, and represent a deliberate operating point rather than a post-hoc tuning on the evaluation data. The consistency of accuracy across all three recording locations (ranging from 99.33% to 99.74%) indicates that the quality filter generalises to different camera viewpoints and traffic conditions without site-specific adjustment.

VI. Conclusion

We have presented a complete track-level pipeline for fine-grained VRU classification from fixed infrastructure cameras, built on a Swin Transformer Tiny classifier fine-tuned on a custom four-class dataset of 82,200 bounding box crops that we collected and annotated, and integrated with YOLO11 person detection, BoT-SORT multi-object tracking, and a confidence gated voting scheme for stable label assignment for each track. Evaluated on 2,965 annotated tracks drawn from three urban surveillance recordings at three distinct camera

locations, the pipeline achieves 99.70% overall accuracy with no systematic inter-class confusion.

We also contribute a clean evaluation protocol in which the pipeline is run once on each evaluation video, tracks are annotated directly, and accuracy is measured by track identifier lookup with no spatial re-matching. This design eliminates a common source of artificial evaluation error and produces a reliable estimate of true classifier performance.

Limitations. Although the evaluation spans three distinct camera locations, it represents a single deployment context in terms of geography, sensor hardware, and resolution. The quality filter restricts the effective operating range to road users within approximately 15–20 m of the camera on 704×576 footage. Detection recall for distant or partially occluded VRUs is not characterised by this evaluation. The training dataset cannot be released in its current form due to GDPR constraints on the source footage, a privacy-anonymized version is under preparation and will be made publicly available alongside the companion benchmark paper.

Future work. We plan to evaluate the pipeline across a wider range of cameras, geographic locations, and lighting conditions, and to investigate the effect of camera angle and resolution on classification performance. We also intend to extend the evaluation to characterise end-to-end detection recall in addition to classifier accuracy, and to explore lighter classifier backbones to reduce inference latency for real-time deployments on embedded hardware.

REFERENCES

[1] World Health Organization, "Global status report on road safety 2023," WHO, Geneva, Tech. Rep., 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240086517

[2] N. Buch, S. A. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 3, pp. 920–939, Sep. 2011.

[3] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4457–4465.

[4] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 3, pp. 413–430, Sep. 2007.

[5] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila, "A new benchmark for vision-based cyclist detection," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, Gothenburg, Sweden, Jun. 2016, pp. 1028–1033.

[6] R. M. Silva, G. F. Azevedo, M. V. V. Berto *et al.*, "Vulnerable road user detection and safety enhancement: A comprehensive survey," *Expert Systems with Applications*, vol. 292, p. 128529, 2025.

[7] G. Wu *et al.*, "Vulnerable road user detection for roadside-assisted safety protection: A comprehensive survey," *Applied Sciences*, vol. 15, no. 7, p. 3797, 2025.

[8] S. Gilroy, D. Mullins, E. Jones, A. Parsi, and M. Glavin, "E-scooter rider detection and classification in dense urban environments," *Results in Engineering*, vol. 16, p. 100677, Dec. 2022.

[9] H. Nguyen, M. Nguyen, and Q. Sun, "Electric scooter and its rider detection framework based on deep learning for supporting scooter-related injury emergency services," in *Innovations in Intelligent Systems and Applications*. Cham: Springer, 2021, pp. 221–232.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10 002.

[11] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," arXiv preprint arXiv:2410.17725, Oct. 2024.

[12] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," arXiv preprint arXiv:2206.14651, Jun. 2022.

[13] K. Saleh, M. Hossny, A. Hossny, and S. Nahavandi, "Cyclist detection in LiDAR scans using faster R-CNN and synthetic depth images," in *Proceedings of the IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Yokohama, Japan, Oct. 2017, pp. 1–6.

[14] Z. Christoforou, C. Gioldasis, Y. Valero, and G. Vasileiou-Voudouris, "Smart traffic data for the analysis of sustainable travel modes," *Sustainability*, vol. 14, no. 18, p. 11150, 2022.

[15] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 3464–3468.

[16] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sep. 2017, pp. 3645–3649.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2021.

[18] Z. Fang and A. M. López, "Intention recognition of pedestrians and cyclists by 2D pose estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4773–4783, Nov. 2020.

[19] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[20] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4913–4922.

[21] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim, "Cannot see the forest for the trees: Aggregating multiple viewpoints to better classify objects in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 052–17 061.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[23] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

[25] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolo," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics