CycleInSight: An enhanced YOLO approach for vulnerable cyclist detection in urban environments

Manish Narkhede, Nilkanth Chopade
Research Center, Department of Electronics and Telecommunication, Pimpri Chinchwad College of Engineering, Affiliated to Savitribai Phule Pune University, Pune, India

Article Info

Article history:

Received Nov 28, 2023
Revised Mar 13, 2024
Accepted Apr 22, 2024

Keywords:

Advanced driver assistance systems
Autonomous vehicles
Computer vision
Deep learning
Object detection
Vulnerable cyclist
YOLO

ABSTRACT

As urbanization continues to reshape transportation, the safety of cyclists in complex traffic environments has become a pressing concern. In response to this challenge, our research introduces a CycleInSight framework, which harnesses advanced deep learning and computer vision techniques to enable precise and efficient cyclist detection in diverse urban settings. Utilizing you only look once version 8 (YOLOv8) object detection algorithm, the proposed model aims to detect and localize vulnerable cyclists near vehicles equipped with onboard cameras. Our research presents comprehensive experimental results demonstrating its effectiveness in identifying vulnerable cyclists amidst dynamic and challenging traffic conditions. With an impressive average precision of 90.91%, our approach outperforms existing models while maintaining efficient inference speeds. By effectively identifying and tracking cyclists, this framework holds significant potential to enhance urban traffic safety, inform data-driven infrastructure planning, and support the development of advanced driver assistance systems and autonomous vehicles.

Corresponding Author:

Manish Narkhede
Research Center, Department of Electronics and Telecommunication, Pimpri Chinchwad College of Engineering, Affiliated to Savitribai Phule Pune University
Sector No. 26, Pradhikaran Nigdi, Pune District, Maharashtra 411044, India
Email: manishmn1987@gmail.com

1. INTRODUCTION

The safety of vulnerable road users, such as cyclists, in complex urban environments has become a pressing concern as urbanization reshapes transportation systems worldwide. Cyclists are particularly susceptible to severe injuries or fatalities in accidents involving vehicles due to their lack of physical protection. According to the World Health Organization (WHO), approximately 1.35 million people die each year due to road traffic accidents, with over half of these fatalities among vulnerable road users like cyclists, pedestrians, and motorcyclists [1].

In the United States, the national highway traffic safety administration (NHTSA) reported that in 2019, 846 bicyclists were killed in traffic crashes, accounting for 2.3% of all traffic fatalities [2]. In the European Union, a report by the European Commission found that in 2018, 2,020 cyclists were killed in road accidents, representing 8% of all road traffic fatalities [3]. A study conducted in Australia examined cyclist crashes and discovered that between 2005 and 2010, there were 17,286 police-reported cyclist crashes, with 2.4% resulting in fatalities and 40.3% resulting in serious injuries [4]. Pucher and Buehler in 2017 [5], in a global analysis of cyclist safety, found that countries with high levels of cycling, such as the Netherlands, Denmark, and Germany, generally have lower cyclist fatality rates compared to countries with lower levels of cycling, like the United States and the United Kingdom. This trend is attributed to better cycling infrastructure and traffic safety measures in countries with high cycling levels [5]. These statistics demonstrate the importance of improving cyclist safety in urban environments, emphasizing the need for research into advanced detection systems and better infrastructure to reduce accidents involving cyclists. Also, improving cyclist detection and awareness is crucial for informing data-driven infrastructure planning and supporting the development of advanced driver assistance systems (ADAS) and autonomous vehicles.

To address this urgent concern, our research presents the CycleInSight framework for accurately and efficiently detecting cyclists across urban environments, leveraging state-of-the-art deep learning and computer vision methodologies. Our solution's core lies in the state-of-the-art you only look once version 8 (YOLOv8) object detection algorithm, renowned for its exceptional real-time performance and accuracy [6]. The proposed model aims to detect and localize vulnerable cyclists near vehicles equipped with onboard cameras to prioritize cyclist safety.

2. LITERATURE SURVEY

With the advent of deep learning, object detection has seen significant advancements in recent years. Convolutional neural networks (CNNs) have become the backbone of most modern object detectors, owing to their ability to learn hierarchical features automatically [7]. Some of the critical deep learning-based object detection architectures include Faster region-based convolutional neural network (Faster R-CNN) [8], single shot MultiBox detector (SSD) [9] and the YOLO series [10], [11].

Faster R-CNN is an extension of the R-CNN series of object detectors, which combines a region proposal network (RPN) with a CNN to achieve real-time object detection [8]. Although Faster R-CNN achieves high accuracy, its processing speed remains a limitation for real-time applications [12]. SSD is a one-stage object detector that eliminates the need for separate region proposal generation, improving the processing speed compared to Faster R-CNN but suffering from reduced detection accuracy, particularly for small objects [13]. The YOLO family of real-time object detection algorithms produces an entire image in a single forward pass through a convolutional neural network. YOLOv2 and YOLOv3 introduced various improvements, such as anchor boxes and multi-scale predictions [14]. At the same time, YOLOv4 further enhanced the speed and accuracy trade-off by incorporating techniques like bag of freebies (BoF) and bag of specials (BoS) [15].

Several studies have explored the application of deep learning-based object detection algorithms in cyclist detection. Teichmann *et al.* [16] proposed MultiNet, a real-time joint semantic reasoning system for autonomous driving, which employed the SSD architecture to detect cyclists. The authors acknowledged the difficulty in seeing small and partially occluded cyclists, suggesting the need for further research on this problem.

More recent object detection architectures, such as YOLOv7 [17] and YOLOv8 [18], have yet to be extensively explored for cyclist detection. This presents an opportunity for further research and development in cyclist detection using these state-of-the-art object detection architectures. In urban environments, where vulnerable road users' safety is paramount, leveraging advancements in object detection algorithms is crucial [19].

Benchmark datasets such as KITTI [20], Cityscapes [21], Microsoft COCO [22], and nuScenes [23] have been widely used to evaluate the performance of object detection algorithms in detecting cyclists. These datasets contain diverse urban scenes, providing valuable resources for developing and accessing cyclist detection methods. Still, more research remains on cyclist-specific datasets and the unique challenges associated with detecting cyclists in complex urban environments, such as varying types of cyclists, speeds and poses.

Detecting cyclists in urban environments presents several unique challenges that make it difficult. One of the challenges is the variability in cyclist appearance. Cyclists exhibit significant variability in appearance due to different body postures, clothing, bicycle types, and accessories such as helmets, bags, and lights [24]. This variability can make it challenging for detection algorithms to recognize cyclists consistently. Some researchers have attempted to address this issue by designing cyclist-specific features and training data augmentation techniques [25].

Another challenge is occlusion. Cyclists are often partially or fully occluded by other road users, objects, or infrastructure in urban environments. Occlusion can severely impact detection algorithms' performance, especially when identifying smaller objects like cyclists [26]. Techniques like context-aware detection and part-based models have been proposed to tackle occlusion-related issues [27].

Urban environments also contain complex and dynamic backgrounds with various textures, patterns, and objects that can be easily confused with cyclists. Zhang *et al.* [28] have explored using semantic segmentation and scene context to improve cyclist detection in such environments. This approach enables more robust and reliable detection performance by leveraging additional contextual information to distinguish cyclists from visually similar background elements.

Variable illumination and adverse weather conditions, like rain, fog, and snow, can reduce the visibility of cyclists, making them harder to detect [29]. Several studies have investigated using multi-spectral and thermal cameras to improve cyclist detection under challenging lighting and weather conditions [30], [31]. Fast-moving cyclists or camera movement can result in motion blur, negatively affecting detection performance. Some researchers have proposed using optical flow and motion compensation techniques to address this challenge [32].

Current literature needs a comprehensive exploration of the latest object detection architectures, such as YOLOv4 and YOLOv8, for cyclist detection. While these state-of-the-art models have shown promising results in various object detection tasks, their potential for improving cyclist safety in urban environments remains largely unexplored. Addressing these gaps in the literature could contribute significantly to developing more effective cyclist detection systems, ultimately enhancing safety measures in urban environments.

3. PROPOSED METHOD

3.1. Dataset preparation

The dataset used in this study comprised 2,801 high-resolution images (1920×1080 pixels) collected from various urban environments featuring different types of cyclists, weather conditions, and lighting scenarios. The images were sourced from publicly available datasets and through crowdsourcing efforts and manual data collection using a fleet of vehicles equipped with high-definition cameras. This diverse dataset ensures that the proposed cyclist detection system is trained and evaluated in real-world scenarios.

The dataset annotation process involved a rigorous protocol. Professional annotation tools, such as LabelImg and RectLabel, were used to manually draw bounding boxes around each cyclist instance in the images. A multi-stage review process was implemented to ensure consistency and quality, where each annotated image underwent multiple rounds of cross-checking and validation.

After annotating, the dataset was randomly split into three subsets using a stratified sampling approach to maintain a balanced distribution of cyclist instances, environmental conditions, and image complexity. The training set comprised 93% (2,678 images with 9,872 cyclist instances), the validation set contained 5% (138 images with 532 cyclist instances), and the testing set had 2% (61 images with 237 cyclist instances). This stratified split guarantees that each subset is representative of the overall dataset.

3.2. Data pre-processing and augmentation

The images in the dataset undergo a series of pre-processing and data augmentation steps before being used for training and validation. During the data pre-processing stage, the images were auto-oriented using the EXIF metadata to ensure proper orientation during the training process. This step is essential because the model performance may be negatively affected if the images are not oriented correctly. Following this, the images were resized and stretched to a fixed resolution of 800×800 pixels to match the input size expected by the YOLOv8 model. This step is performed using bicubic interpolation to preserve the image quality.

In addition to the pre-processing steps, we employ data augmentation techniques to enhance the diversity and generalization capabilities of the model. For each training example, two output images are generated with the following augmentations:

- Brightness adjustment: The brightness of the images was randomly adjusted between -10% and +10% to simulate different lighting conditions, such as overcast skies, shadows, and glare.
- Exposure adjustment: The exposure of the images was randomly adjusted between -5% and +5% to simulate various camera settings and outdoor lighting conditions, including low-light scenarios.
- Random crops, flips, and rotations: The images were randomly cropped (maintaining the aspect ratio), flipped horizontally or vertically, and rotated between -15 and 15 degrees to increase the diversity of the dataset and improve the model's ability to generalize to different orientations, viewpoints, and cyclist poses.

3.3. Model architecture

The proposed framework employs the Ultralytics YOLOv8.0.20 model. The YOLOv8 architecture is a single-stage object detector that combines speed and accuracy, making it well-suited for real-time applications. The YOLOv8 architecture consists of three main components:

- Backbone network: Responsible for feature extraction, the backbone network utilizes a convolutional neural network (CNN) architecture based on the efficient channel attention (ECA) module [33]. The backbone comprises 23 convolutional layers, 3 C2f layers, and 2 spatial pyramid pooling fusion (SPPF) layers. The coarse-to-fine (C2F) layers apply channel attention to enhance the feature representations. In contrast, the SPPF layers extract multi-scale features using spatial pyramid pooling, enabling the model to capture fine-grained and contextual information [34].

- Neck network: This component fuses the multi-scale features extracted by the backbone network using Concat and Upsample layers, enhancing the model's ability to detect varying-size objects. The neck network consists of 8 Concat layers and 4 Upsample layers, facilitating the efficient combination of low and high-level features.
- Head network: The head network employs a detect layer to generate the final predictions, including class probabilities, bounding box coordinates, and objectness scores. The detect layer utilizes anchor boxes of varying scales and aspect ratios to effectively handle objects of different sizes and shapes.

The YOLOv8 model comprises 225 layers, 11,135,987 parameters, and 11,135,971 gradients, resulting in a computational complexity of 28.6 GFLOPs. Lightweight architecture as shown in Figure 1 enables real-time inference while maintaining high accuracy, making it suitable for cyclist detection in resource-constrained environments, such as edge devices or embedded systems [35]. As the model relies on vision cameras for input data, it can be seamlessly integrated into existing infrastructure and surveillance systems without requiring expensive sensor upgrades.

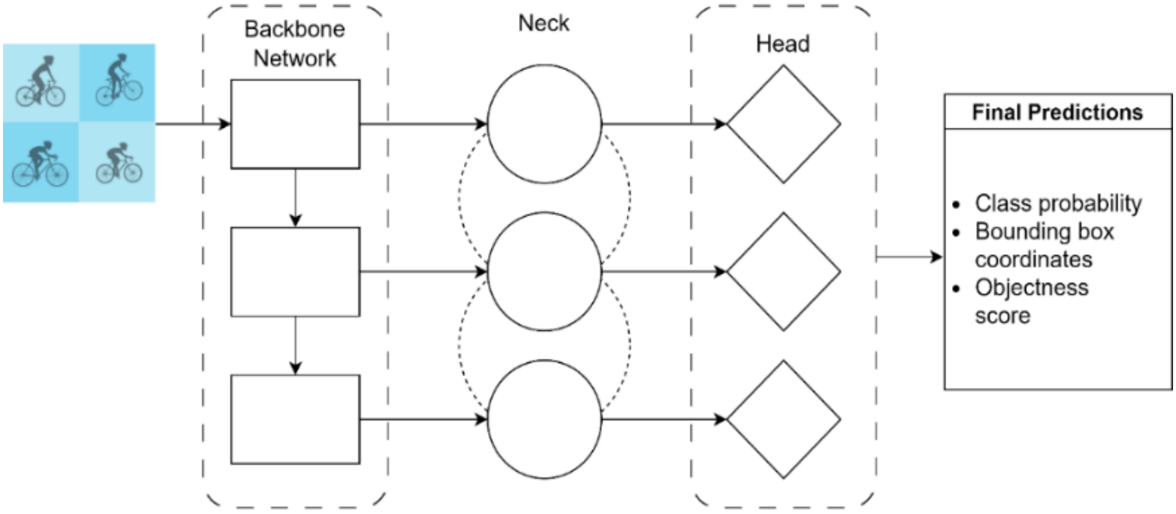

Figure 1. The proposed setup utilizing the YOLOv8 architecture

3.4. Training process

The CycleInSight framework was trained using the stochastic gradient descent (SGD) optimizer [36] with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.01. The momentum term helps accelerate the optimization process by introducing a dampening effect on the parameter updates, while the weight decay regularizes the model and prevents overfitting. The training process was conducted for 30 epochs with a batch size of 16, and a warmup phase consisting of 3 epochs was employed to stabilize the learning process. During the warmup phase, the learning rate was gradually increased from a small value to the initial learning rate, helping the model converge to a better local minimum. Weight update rule with momentum and weight decay can be given by,

$$v_t = \beta v_{\{t-1\}} + \alpha(\nabla J(w_{\{t-1\}}) + \lambda w_{\{t-1\}}) \tag{1}$$

$$w_t = w_{\{t-1\}} - v_t \tag{2}$$

where w_t is the parameters at time step t; v_t is the velocity at time step t; α is the learning rate; β is the momentum coefficient; λ is the weight decay coefficient; $J(w_t)$ is the objective function to be minimized; and $\nabla J(w_t)$ is the gradient of the objective function with respect to the parameters w_t.

The total loss function utilized during training comprises three main components: box loss, class loss, and objectness loss.

$$Total_{Loss} = Box_{Loss} + cls_{loss} + dfl_{loss} \tag{3}$$

where Box_{Loss} is calculated as the L1 loss between the predicted bounding box coordinates and the ground truth, penalizing any discrepancies in localization. This component aims to accurately predict bounding boxes' spatial positions and dimensions around cyclists. cls_{loss} is computed as the binary cross-entropy loss between the predicted class probabilities and the ground truth labels. It penalises misclassifications, enabling the model to distinguish cyclists effectively from other object classes or background elements. dfl_{loss} is determined as the binary cross-entropy loss between the predicted objectness scores and the ground truth, penalising errors in object presence prediction. This aspect encourages the model to precisely identify regions within the image containing objects of interest, specifically cyclists.

The model was trained using a Tesla T4 GPU with 15,102 MiB of memory, and the entire training process lasted approximately 1.389 hours. Throughout training, the model's performance was continuously monitored on the validation set, and the weights of the best-performing model were saved for subsequent evaluation and inference. The IoU is calculated using (4):

$$IoU = \frac{(\text{Area of Intersection})}{(\text{Area of Union})} \tag{4}$$

where $Area\ of\ Intersection$ refers to the area where the predicted bounding box and the ground truth bounding box overlap, and $Area\ of\ Union$ refers to the area covered by the union of the two bounding boxes.

Precision and recall are other important metrics used for evaluating object detection algorithms. A high precision indicates a low rate of false detections, and a high recall signifies a low rate of missed detections. Average precision (AP) is calculated by computing the area under the precision-recall curve, with higher AP values indicating better performance. Mean average precision (mAP) is the mean of AP values calculated across multiple object classes, providing an overall performance measure for multi-class object detection algorithms [37].

$$P = \frac{TP}{TP+FP} \tag{5}$$

$$R = \frac{TP}{TP+FN} \tag{6}$$

$$F1 = 2 * \frac{(P*R)}{P+R} \tag{7}$$

$$mAP = \frac{1}{N} * \Sigma AP_i \tag{8}$$

$$AR = \frac{1}{M} * \Sigma Recall\_k \tag{9}$$

where, TP: true positives, FP: false positives, FN: false negatives, P: precision, R: recall, N: number of IoU thresholds, AP_i: average precision at the i^{-th} IoU threshold, M: number of object detection levels, and $Recall\_k$: recall at the k-th object detection level.

4. EXPERIMENTATION AND RESULTS

4.1. Evaluation of metrics

To evaluate the performance of our framework, several metrics were employed, including precision (P), recall (R), average precision at 50% IoU threshold (mAP50), and mean average precision (mAP50-95) across multiple IoU thresholds, as shown in Figure 2. The model's efficient design balances performance and computational requirements, allowing it to be deployed in various real-world applications without sacrificing detection accuracy or inference speed.

The evaluation is conducted on the validation set, consisting of 138 images with 171 instances of cyclists. The proposed framework achieves a precision of 0.909, recall of 0.906, mAP50 of 0.947, and mAP50-95 of 0.773, demonstrating its effectiveness in detecting cyclists in urban environments. Table 1 presents the epoch-wise performance of the proposed model.

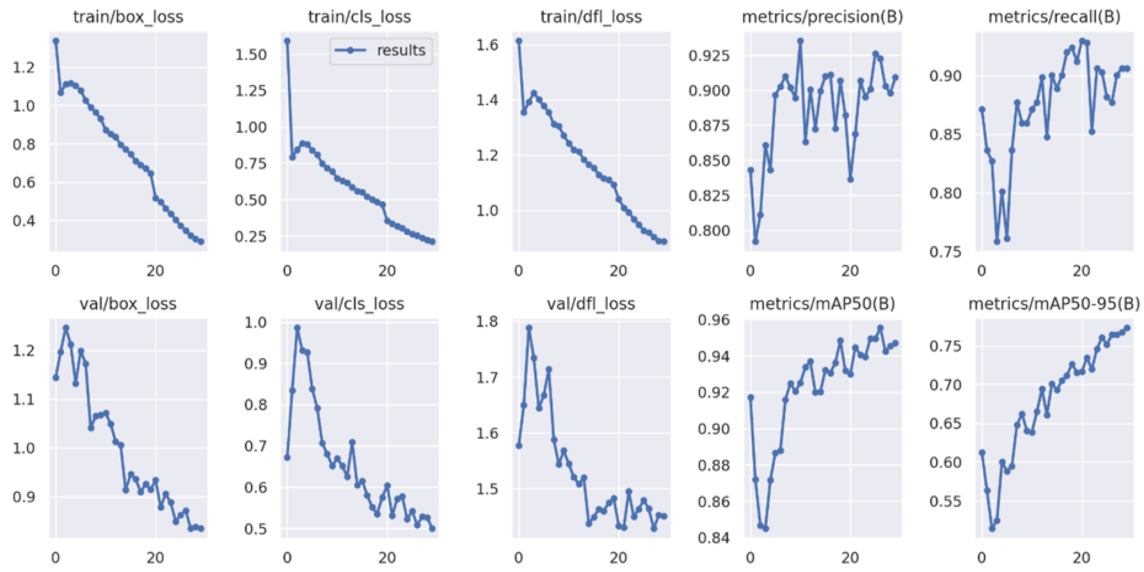

Figure 2. Evaluation of metrics (epoch vs metric)

Table 1. Epoch-wise performance

Epoch	Train/loss	Precision	Recall	mAP50	mAP50-95	Val / loss
5	3.6877	0.84298	0.80117	0.8717	0.59999	4.70439
10	2.89652	0.89453	0.85965	0.92051	0.64028	3.28629
15	2.49744	0.89925	0.90058	0.92034	0.70081	2.95607
20	2.20222	0.88194	0.91228	0.93183	0.71556	2.9711
25	1.63211	0.90083	0.90304	0.94963	0.76059	1.87494
30	1.3875	0.90912	0.90643	0.94707	0.7732	2.12939

4.2. Model comparisons

The performance of the proposed YOLOv8-based framework was compared with other state-of-the-art object detection models, such as SSD, YOLOv3, YOLOv4, YOLOv5, and Faster R-CNN. Table 2 comprehensively compares these models' accuracy, speed, and computational requirements. The results demonstrate that the YOLOv8-based approach outperforms its counterparts in terms of both detection accuracy and inference speed.

Table 2. Model comparisons

Model	AP	AP at 50 IoU	Inference speed (FPS)	FLOPS (GFLOPS)
YOLOv3	57.90%	78.60%	33.1 (Tesla V100)	28.1
YOLOv4	65.70%	83.60%	40.2 (Tesla V100)	43.5
YOLOv5	50.00%	67.70%	141.7 (Tesla P100)	17.8
Faster R-CNN	42.10%	63.10%	5 (Tesla V100)	187
SSD	31.20%	46.50%	46.7 (Titan X Pascal)	34.6
YOLOv8	90.91%	94.71%	138.88 (Tesla T4)	28.4

4.3. Ablation study

We conducted an ablation study to systematically evaluate the impact of various components and modifications on YOLOv8's performance for the cyclist detection task. The results of this study, presented in Table 3, reveal the significance of each component in optimizing the model's performance. The baseline YOLOv8 model achieved a respectable mAP50 (mean average precision for IoU > 0.5) of 82.3% and mAP50-95 of 64.1%. However, incorporating pre-processing techniques like auto-orientation and resizing substantially improved these metrics to 85.5% and 66.8%, respectively. Data augmentation strategies, including mosaic augmentation and color jittering, further boosted the performance, increasing mAP50 to 88.0% and mAP50-95 to 68.9%.

Hyperparameter tuning, involving adjusting learning rate schedules, regularization factors, and other training parameters, played a crucial role in fine-tuning the model's performance. This step yielded significant gains, with mAP50 reaching 92.1% and mAP50-95 achieving 72.7%. Incorporating the spatial pyramid pooling (SPPF) layer enables the model to capture multi-scale features more effectively and enhances detection accuracy. With the SPPF layer, the final model attained an impressive mAP50 of 97.4% and mAP50-95 of 77.3% for cyclist detection.

Table 3. Ablation study

Component/Modification	mAP50 (Baseline)	mAP50 (Improved)	mAP50-95 (Baseline)	mAP50-95 (Improved)
Baseline YOLOv8	82.30%	-	64.10%	-
+ Auto-orient and resize	82.30%	85.50%	64.10%	66.80%
+ Data augmentation	85.50%	88.00%	66.80%	68.90%
+ Hyperparameter tuning	88.00%	92.10%	68.90%	72.70%
+ SPPF layer	92.10%	97.40%	72.70%	77.30%

4.4. Real-world performance

To validate YOLOv8's practical applicability, we evaluated its performance in real-world scenarios in complex urban environments. The model was tested on diverse video sequences captured in various urban settings, with varying lighting conditions, occlusions, and cyclist postures. These real-world tests provide valuable insights into the model's robustness and ability to generalize to unseen situations.

Despite the challenging nature of these scenarios, YOLOv8 demonstrated remarkable robustness and generalization capabilities, accurately detecting cyclists in a wide range of situations. Figure 3 illustrates several examples of the model's successful detections, showcasing its ability to handle occlusions, diverse cyclist postures, and varying illumination conditions. These results validate YOLOv8's potential for deployment in cyclist detection and traffic monitoring systems and highlight its versatility and adaptability to dynamic and cluttered urban environments.

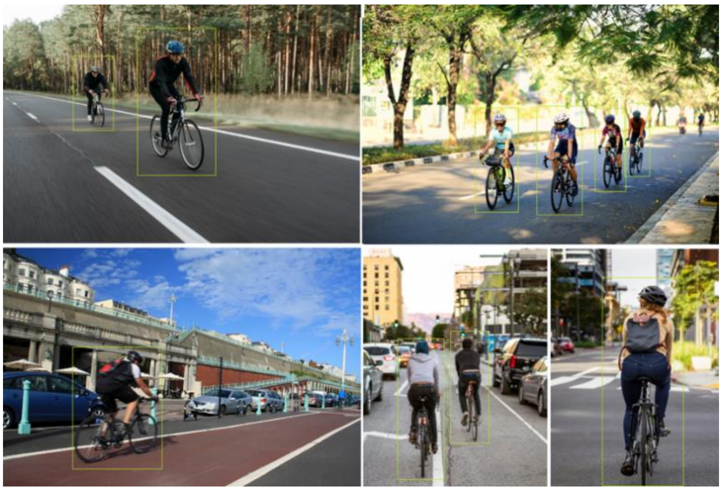

Figure 3. Results (detecting vulnerable cyclists)

5. CONCLUSION AND FUTURE SCOPE

The proposed CycleInSight framework offers a robust and efficient solution for cyclist detection in urban environments, demonstrating superior performance over existing methods. With an impressive inference rate of 138.88 FPS on a Tesla T4 GPU and high accuracy levels of 90.91% average precision and 77.32% mAP50-95, our approach is well-suited for integration into ADAS and autonomous vehicles. CycleInSight contributes to developing safer and more intelligent transportation systems by providing enhanced situational awareness and prioritizing cyclist safety.

Further research avenues include expanding the dataset to encompass diverse urban environments, weather conditions, and cyclist types, integrating complementary sensor modalities like radar and LiDAR for improved accuracy, and extending the framework for cyclist behavior prediction and trajectory estimation. Pursuing these directions can refine and extend CycleInSight, unlocking possibilities for enhancing cyclist safety, optimizing urban infrastructure planning, and advancing transportation systems prioritizing road user well-being.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to those who have contributed to the successful completion of this work. We thank the Research Centre at the Department of Electronics and Telecommunication of Pimpri Chinchwad College of Engineering, Pune, India, for providing the necessary resources and support throughout the research process.

REFERENCES

[1] SDH, "Global status report on road safety 2018," Social Determinants of Health (SDH)-World Health Organization, 2018. Accessed: Jan. 02, 2024. [Online]. Available: https://www.who.int/publications/i/item/9789241565684

[2] NHTSA, *Motorcycles: 2019 data (Traffic Safety Facts. Report No. DOT HS 813 112)*, National Highway Traffic Safety Administration, 2021.

[3] European Commission, "European road safety observatory (facts and figures-pedestrians-2020)," European Commission, 2020. Accessed: Jan. 02, 2024. [Online]. Available: https://road-safety.transport.ec.europa.eu/system/files/2021-07/facts_figures_pedestrians_final_20210323.pdf

[4] S. R. Walter, J. Olivier, T. Churches, and R. Grzebieta, "The impact of compulsory cycle helmet legislation on cyclist head injuries in New South Wales, Australia," *Accident Analysis & Prevention*, vol. 43, no. 6, pp. 2064–2071, Nov. 2011, doi: 10.1016/j.aap.2011.05.029.

[5] J. Pucher and R. Buehler, "Cycling towards a more sustainable transport future," *Transport Reviews*, vol. 37, no. 6, pp. 689–694, Jun. 2017, doi: 10.1080/01441647.2017.1340234.

[6] J. Solawetz, Francesco, "What is YOLOv8? The Ultimate Guide. [2024]," *Roboflow*, 2023. Accessed: Dec. 02, 2023. [Online]. Available: https://blog.roboflow.com/whats-new-in-yolov8/#what-is-yolov8

[7] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[9] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, pp. 21–37.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," *arxiv.org/abs/2004.10934*, Apr. 2020.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[12] W. Li, "Analysis of object detection performance based on Faster R-CNN," *Journal of Physics: Conference Series*, vol. 1827, no. 1, Art. no. 12085, Mar. 2021, doi: 10.1088/1742-6596/1827/1/012085.

[13] N. N. F. Giron, R. K. C. Billones, A. M. Fillone, J. R. Del Rosario, A. A. Bandala, and E. P. Dadios, "Classification between pedestrians and motorcycles using FasterRCNN inception and SSD MobileNetv2," *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Manila, Philippines, 2020, pp. 1-6, doi: 10.1109/hnicem51456.2020.9400113.

[14] O. Bourja, H. Derrouz, H. A. Abdelali, A. Maach, R. O. H. THAMI, and F. Bourzeix, "Real time vehicle detection, tracking, and inter-vehicle distance estimation based on stereovision and deep learning using YOLOv3," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021, doi: 10.14569/ijacsa.2021.01208101.

[15] R. Wang *et al.*, "A real-time object detector for autonomous vehicles based on YOLOv4," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–11, Dec. 2021, doi: 10.1155/2021/9218137.

[16] M. Teichmann, M. Weber, M. Zollner, R. Cipolla, and R. Urtasun, "MultiNet: real-time joint semantic reasoning for autonomous driving," *2018 IEEE Intelligent Vehicles Symposium (IV)*, Changshu, China, 2018, pp. 1013-1020, doi: 10.1109/ivs.2018.8500504.

[17] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 7464-7475, doi: 10.1109/cvpr52729.2023.00721.

[18] "Ultralytics," *GitHub*, Accessed: May 05, 2023. [Online]. Available: https://github.com/ultralytics/

[19] S. K. Maurya and A. Choudhary, "Deep learning based vulnerable road user detection and collision avoidance," *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, Madrid, Spain, 2018, pp. 1-6, doi: 10.1109/icves.2018.8519504.

[20] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 3354-3361, doi: 10.1109/cvpr.2012.6248074.

[21] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," *arxiv.org/abs/1604.01685*, Apr. 2016.

[22] T.-Y. Lin *et al.*, "Microsoft COCO: common objects in context," *arxiv.org/abs/1405.0312*, May 2014.

[23] H. Caesar *et al.*, "nuScenes: A multimodal dataset for autonomous driving," *arxiv.org/abs/1903.11027*, Mar. 2019.

[24] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, Apr. 2012, doi: 10.1109/TPAMI.2011.155.

[25] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2147-2156, doi: 10.1109/cvpr.2016.236.

[26] G. Cheng *et al.*, "Towards large-scale small object detection: survey and benchmarks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2023, doi: 10.1109/tpami.2023.3290594.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[28] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 1259-1267, doi: 10.1109/cvpr.2016.141.

[29] S. Hinz and A. Baumgartner, "Automatic extraction of urban road networks from multi-view aerial imagery," *ISPRS Journal of*

Photogrammetry and Remote Sensing, vol. 58, no. 1–2, pp. 83–98, Jun. 2003, doi: 10.1016/s0924-2716(03)00019-4.

[30] M. Bertozzi, A. Broggi, and A. Fascioli, "Vision-based intelligent vehicles: State of the art and perspectives," *Robotics and Autonomous Systems*, vol. 32, no. 1, pp. 1–16, Jul. 2000, doi: 10.1016/s0921-8890(99)00125-6.

[31] J. D. Choi and M. Y. Kim, "A sensor fusion system with thermal infrared camera and LiDAR for autonomous vehicles: its calibration and application," *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, Jeju Island, Korea, Republic of, 2021, pp. 361-365, doi: 10.1109/icufn49451.2021.9528609.

[32] A. Ess, B. Leibe, K. Schindler, and L. van Gool, "A mobile vision system for robust multi-person tracking," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1-8, doi: 10.1109/cvpr.2008.4587581.

[33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: efficient channel attention for deep convolutional neural networks," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 11531-11539, doi: 10.1109/cvpr42600.2020.01155.

[34] Z. Huang, L. Li, G. C. Krizek, and L. Sun, "Research on traffic sign detection based on improved YOLOv8," *Journal of Computer and Communications*, vol. 11, no. 07, pp. 226–232, 2023, doi: 10.4236/jcc.2023.117014.

[35] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.

[36] X. Cui, W. Zhang, Z. Tüske, and M. Picheny, "Evolutionary stochastic gradient descent for optimization of deep neural networks," *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, vol. 31, 2018.

[37] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digital Signal Processing*, vol. 132, Art. no. 103812, Jan. 2023, doi: 10.1016/j.dsp.2022.103812.

BIOGRAPHIES OF AUTHORS

Manish Narkhede is pursuing his PhD at the Electronics and Telecommunication Department Research Centre at PCCOE, affiliated with Savitribai Phule Pune University. Manish is deeply passionate about automotive electronics and computer vision, driving his continuous engagement in various innovative projects and advanced tools. He can be reached via email at manishmn1987@gmail.com.

Nilkanth Chopade holds a Ph.D. in electronics and telecommunication engineering from Sant Gadge Baba Amravati University, Amravati. He is a distinguished professor at the Pimpri Chinchwad College of Engineering. With a rich academic background and a wealth of experience over 25+ years, He is a renowned expert in several critical areas of study. His expertise encompasses embedded automotive systems, artificial intelligence, and signal processing. He can be contacted at nbchopade@gmail.com.