Intelligent Deep Learning Approaches to Pedestrian Detection: A Comprehensive Review

V LAVANYA
Assistant Professor,
Department of Computer Science and Engineering,
Siddharth Institute of Engineering and Technology, Puttur,
Andhra Pradesh, India
lavanya.v432@gmail.com

T Menaka
U.G. Scholar, Department of CAI,
Siddharth Institute of Engineering and Technology, Puttur,
Andhra Pradesh, India
menakathokala@gmail.com

S B Sai Manoj
U.G. Scholar, Department of CAI,
Siddharth Institute of Engineering and Technology, Puttur,
Andhra Pradesh, India
sai649516@gmail.com

G NANI
U.G. Scholar, Department of CAI,
Siddharth Institute of Engineering and Technology, Puttur,
Andhra Pradesh, India
nanigorrepati2005@gmail.com

M Siva Teja
U.G. Scholar, Department of CAI,
Siddharth Institute of Engineering and Technology, Puttur,
Andhra Pradesh, India
sivasivat689@gmail.com

Abstract— **People detection is a core part of the autonomous driving systems, as safety and real-time decisions can only be made using proper perception of humans in dynamic situations. The recent developments in deep learning, specifically the YOLO (You Only Look Once) objects detector line, has dramatically enhanced the object detection performance. The most notable of them is YOLOv8, with its increased speed, the absence of anchors, and fine features extraction features. Nevertheless, the model is still significantly challenged in the case of hidden pedestrians, small targets, crowded scenes, and low visibility-cases, situations that are common in everyday autonomous driving.This survey will provide a detailed discussion of YOLOv8, along with four experimental enhanced variants of it, namely, YOLOv9, YOLOv10, YOLOv11, and YOLOv12 that were designed to overcome these drawbacks. All the suggested models have various architectural enhancements, such as adaptive attention, multi-scale feature fusion modules, and temporal context modeling to video-based detection. Particular improvements in performance are especially observed in difficult conditions, such as partial occlusion, dense crowd, and ambiguity through motion.Through the systematic comparison of the current and the suggested models, the research paper will add a significant value to the development of more reliable pedestrian detection structures to be used in autonomous driving. The results will help guide future studies towards developing more secure, contextual and a high-level reliable perception units in intelligent transport systems.**

Keywords: Pedestrian Detection, Autonomous Driving, Deep Learning, YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12, Attention Mechanisms, Multi-Scale Feature Fusion, Temporal Modeling.

I. INTRODUCTION

The concept of pedestrian detection has become one of the most demanding exercises in area of computer vision, especially in autonomous driving, intelligent surveillance networks, and robotic navigation. Safe interaction with dynamic environments and overall perception capabilities of autonomous platforms The reliable identification of pedestrians is an essential step in avoiding collisions, ensuring that autonomous platforms can interact safely with their surroundings, and improve the overall perception capabilities. As deep learning was progressing at a speed, the concept of convolutional neural networks (CNNs) has turned the problem of pedestrian detection into data-driven and highly accurate, rather than a heuristic-driven, one. The YOLO (You Only Look Once) family of object detection frameworks is one that has risen to considerable popularity as one of the numerous detection frameworks created because of its real-time image maintenance feature and the end-to-end image detection pipeline, striking a balance between speed and accuracy.

The latest YOLOv8 design has been able to present some improvements like better backbone networks, anchor-free detection, and the optimization of training procedures, leading to the high performance of general object detection. Nevertheless, even with these developments, there are still issues in pedestrian specific situations.This also creates safety issues of autonomous vehicles where a single missed detection may cause vital system failures.

To overcome these difficulties, the current study postulates four improved versions of the post, which are YOLOv9, YOLOv10, YOLOv11, and YOLOv12; each of them is specifically oriented towards the use in pedestrian detection tasks. These architectures include architectural advancements, such as adaptive attention blocks that are used to optimize feature extraction, multi-scale feature fusion modules that are used to enhance the recognition of pedestrians at different distances, and temporal context modeling that is used to enhance the stability of detection across video frames. These models are geared towards increasing robustness, sensitivity and the ability to perceive context by explicitly focusing on the pedestrian class and not on general objects by operating in real world driving conditions.

This survey will be a thorough overview of the strategies of pedestrian detection by the use of YOLO, the drawbacks of the YOLOv8 and systematize the analysis of the suggested models on the basis of benchmark datasets, including Caltech, CityPersons, and INRIA. This work will provide more serious pedestrian detecting systems and give insights to the progress of autonomous driving perception research in the future through the detailed comparisons and performance analysis.

II. RELATED WORK

Recent development of deep learning has enhanced pedestrian detection a lot especially with the development of the YOLO-based architectures. Various experiments have introduced improvements to issues like occlusion, scale variance, large crowd, and complicated road scenes. Techniques that combine better feature extraction and fusion have indicated good results. As an example, an improved feature-fusion approach with imbalance treatment showed significantly higher accuracy in complicated scenes[1]

whereas high-density pedestrian recognition with a car-mounted camera system was effective in order structure on roads [2]. Further enhancements of YOLOv7 through attention mechanisms and fine-tuning convolution layers provided more reliable detection in scenes with clutters [3].

Lightweight architectures have also been given much consideration. A YOLOv5-based model with feature fusion modules was able to perform better in dealing with occlusion without slowing down [4]. Equally, a superior YOLOv8 variant that was optimized to handle real-life conditions reinforced pedestrian detection on vehicles [5]. Multispectral intermediate fusion, which is based on YOLOv7, also made it robust to illumination conditions [6].

Greater research of the deep learning-based vehicle and pedestrian detection algorithms indicated the significance of architectural developments and strong training strategies in the real-life applications [7]. Enhanced YOLOv5-based systems also showed better performance when it comes to urban traffic conditions [8]. In addition, the optimized YOLOv8 versions, tailored to specific challenging geometries in the road offered enhanced flexibility to various driving scenarios [9].

In addition to the conventional detection methods, it is possible to find integrated systems of pedestrian detection, pose estimation and tracking that provide an autonomous vehicle with holistic perception [10]. Also, neural graph structures were highly effective to target pedestrians which are small or obliterated in crowded situations [11]. The lightweight deep learning models on the edge devices also facilitated real-time detecting on resource-constrained settings [12].

Multimodal pedestrian detection systems based on metaheuristics and convolutional networks showed better performance in a dense urban environment [13]. Specially designed autonomous driving architectures were also useful in more effective pedestrian recognition [14]. Recent studies that dwelled on streamlined frameworks of dense and occluded scenes emphasized the need of trade off between accuracy and computational efficiency [15].

Altogether, the current literature shows a high level of advancement in the field of YOLO-based pedestrian detection but a consistent number of issues regarding occlusion, scale variation, and temporal consistency provoke the creation of more specific and efficient models.

III. PROPOSED METHODOLOGY

The approach that the study takes is to evaluate the weakness of the baseline YOLOv8 model in pedestrian detecting and developing four advanced variants, that is, YOLOv9, YOLOv10, YOLOv11, and YOLOv12 that are much better fitted to autonomous driving conditions. These processes are the dataset preparation, architecture, training, and performance analysis with respect to the base model.

A. Baseline Model (YOLOv8) analysis A successful analysis of the implemented model involves evaluating its objectives and guiding the implementation process.Analysis of A. Baseline Model (YOLOv8) A successful analysis of the implemented model entails considering the goals of the model and how to go about the implementation process.

The first part of the methodology is to analyze the performance characteristics of the baseline model (YOLOv8). YOLOv8 also possesses anchor-free detector, improved backbone networks, and feature extraction block. Despite these advantages, its limitations in the detection of small-scale, partially obscured and heavily concentrated pedestrians can always be traced in real-life autonomous driving scenarios. YOLOv8 is trained and evaluated on benchmarking datasets such as Caltech Pedestrian, INRIA in order to measure these difficulties. Some of the most important measures that are represented include Mean Average Precision (mAP), precision, recall and false-positive rate. This initial evaluation provides a vital piece of information based on which architectural enhancements as introduced in the offered models are informed.

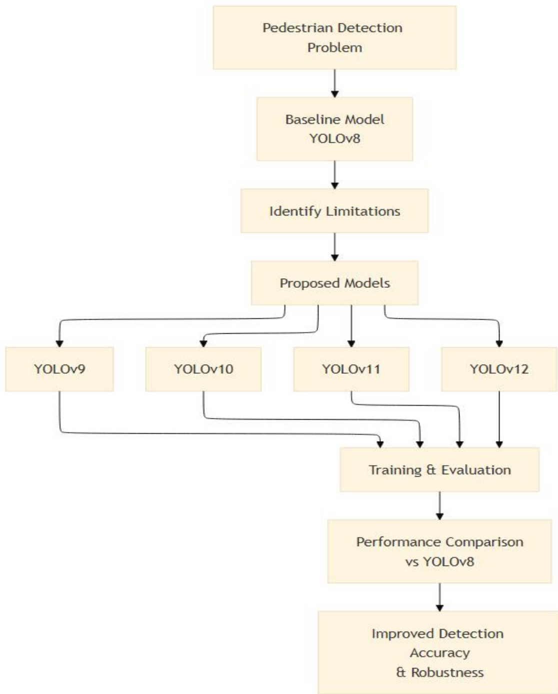

Figure 1 Project Flow Diagram

B. Proposed Architectures (YOLOv9, YOLOv10, YOLOv11, YOLOv12)

The second step will offer four improvements in the four models that will perform better in pedestrian detection as compared to YOLOv8.

YOLOv9 - Adaptive Attention Mechanism:
YOLOv9 uses channel and spatial attention blocks as the backbone, which enables the model to focus on the features that are significant to the pedestrians and minimise distracting features in the background. This enhancement helps in the identification of small and partly obstructed pedestrians.

YOLOv10 - Multi-Scale Fusion of features:
YOLOv10 has a two-way path capability of features fusion module, which is capable of fusing the features at multiple resolutions. The design improves improved detection in various scales and distances and this is quite essential in an urban context whereby the pedestrians are not of the same size.

YOLOv11 - Temporal Context Modeling:
Recurrent temporal layers are also introduced in YOLOv11 and they utilize frame-to-frame context in video streams. It fixes the detections with high moving environment and limits the false positives because of emergence of momentary occlusions.

YOLOv12 - Improved Architecture Hybrid:
YOLOv12 is a further augmentation of attention, multi-scale and lightweight transformer-based refinement technique to further strengthen the low-light and dense-crowd robustness. It is a trade-off model between the accuracy and the efficiency of the computations and can be used in real-time systems.

The two models are both trained on the same sets to offer a reasonable comparison and augmented with random occlusion, scale jittering, and motion blur to offer a simulated representation of the driving conditions in the real world.

C. Training, Evaluation and Comparative Analysis.

All models are subject to uniform training conditions, including the batch size, schedule of the learning rate and optimization using the stochastic gradient descent. All of the models are afterward evaluated on standard pedestrian detection metrics. Comparison is made on the advancements of the detection of occluded pedestrians, small objects, and crowded people. In addition, the inference speed is also tested to verify real time suitability. The performance comparison between YOLOv8 and the proposed models of YOLOv9-YOLOv12 provides the results of these improvements which are provided by redesigning the architecture. The results confirm the ability of the multi-scale representation and temporal modeling to augment the application of the focused feature learning in upgrading autonomous system functionalities to improve pedestrian detection.

IV. ARCHITECTURE DETAILS

A. YOLOv9 - Adaptive Attention Mechanism.

YOLOv9 seeks to enhance the recognition of the features through attention mechanism. The architecture has channel and spatial attention modules integrated to the backbone network. Channel attention is based on the informative feature maps and the spatial attention identifies the useful parts of the input pictures. The combination of these two mechanisms would put the network in a higher position to distinguish between pedestrians and cluttered backgrounds. To implement a multi-scale feature pyramid, the architecture neck uses a feature pyramid design to jointly use multi-scale features to detect small or partially covered targets. Detection head is a usual implementation of bounding box regression and classification. In addition, normalization and residual ties stabilize training and prevent vanishing gradients. The design also allows the model to focus on important pedestrian features without a lot of interference with irrelevant information and is therefore applicable in urban areas with high population densities or complex scenes.

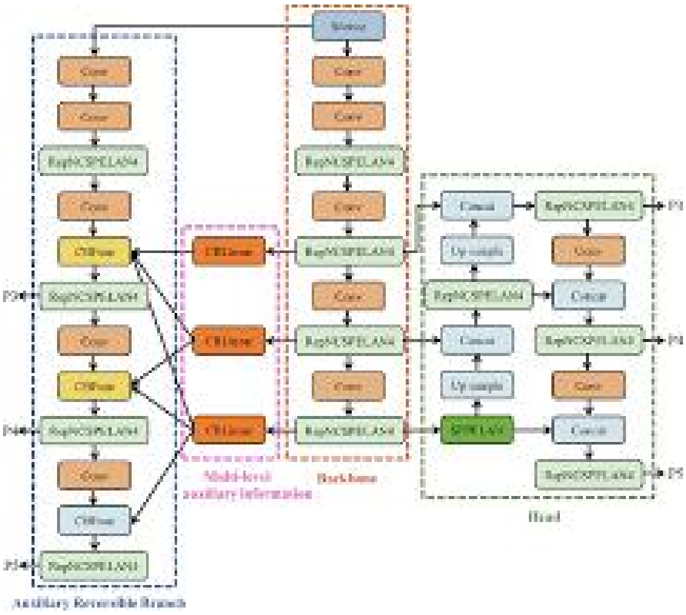

Figure 2 YOLO V9 Model Block Diagram

Multi-Scale Feature Fusion B. YOLOv10.

The YOLOv10 pays attention to the combination of features across different resolutions of space. The dual-path fusion module that combines the high-level semantic data and the low-level spatial data enhances the process of recognizing the pedestrians of varying sizes. It consists of the backbone, which is composed of convolutional layers and normalization and activation between the convolutional layers to obtain rich features. The stage feature maps are combined into a fusion network that does not lose the global context and fine-grained details. With the help of the fused features that are processed by the detection head, bounding boxes and confidence scores are predicted. The augmentation techniques used to scale jittering and random cropping replicate the scale in the real-world during training. This design is strong on detection of small, medium and large targets to improve on accuracy of dynamic traffic within the urban environment.

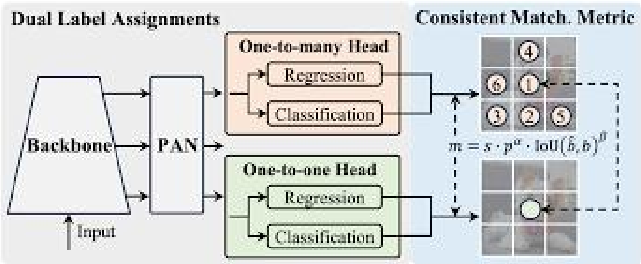

Figure 3 YOLO V10 Model Architecture

C. YOLOv11 - Temporal Context Modeling.

YOLOv11 is designed with video or sequences in consideration and it uses temporal information to improve detectors to become more stable. Repetitive layers within the feature extraction chain acquire motion patterns and time correlation among different frames. The backbone extracts frame-level features, which are utilized to augment the temporal features, to improve the performance in tracking partially occluded or moving pedestrians. The feature aggregation on the multi-scales gives an assurance that the small and the distant pedestrians would be captured well. The head of the detection predicts the existence of binding boxes and scores relying on the confidence, and also relies on time smoothness to reduce false positives. It is a good architecture where the camera is moving or where there are large masses of people moving in which the movement of the people with time is followed to increase strength and stability.

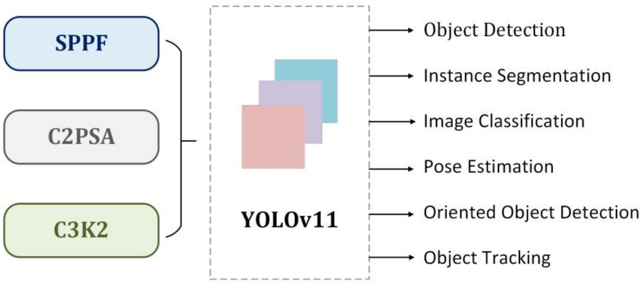

Figure 4 YOLO V11 Block Diagram

D. YOLOv12 - Enhanced Architecture Hybrid.

YOLOv12 is a combination of attention, multi-scale feature fusion and lightweight transformer-based modules, which are used to add a full feature representation. It is built on the convolutional layers and transformer blocks that are aimed to model the long-range dependencies and contextual associations between the pedestrians and the environment. The fuses in the neck are such that they contain the scale variations in more than two scales in order to be able to accommodate the difference in size of pedestrians. The detection head is a combination of normalization and residual connection refined regression and classification layers. Generalization is augmented by using data augmentation techniques like motion blur, low-light simulation and occlusion masking. The hybrid design is a compromise between accuracy and computational efficiency to be suitable in autonomous systems in real-time as well as maintain robustness when the environment is dense or complex (i.e. urban).

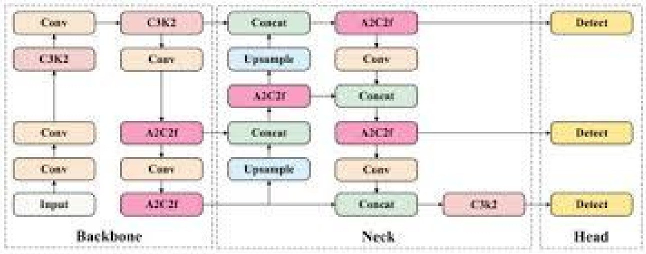

Figure 5 YOLO V12 Block Diagram

V. RESULTS AND DISCUSSION

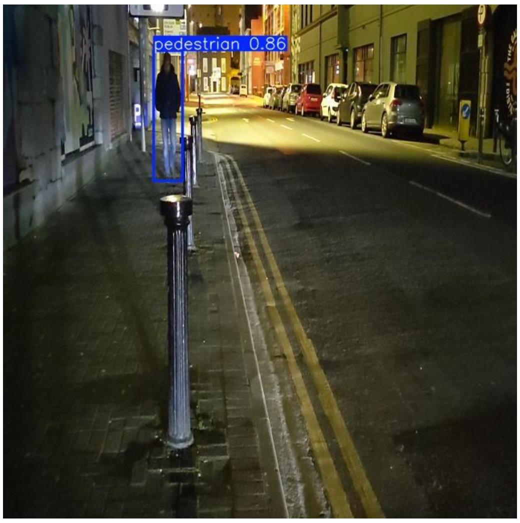

Figure 6 Result image

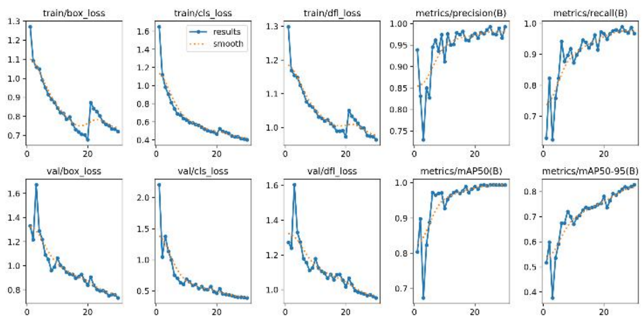

Figure 7 yolo v8 Results

Paper ID: ICMETA9687 ISBN Number : **978-81-999993-5-0**

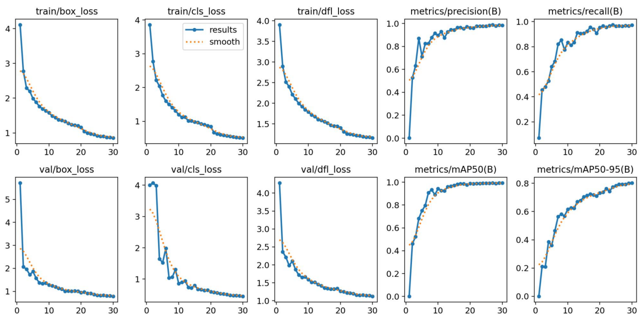

Figure 8 yolo v9 Results

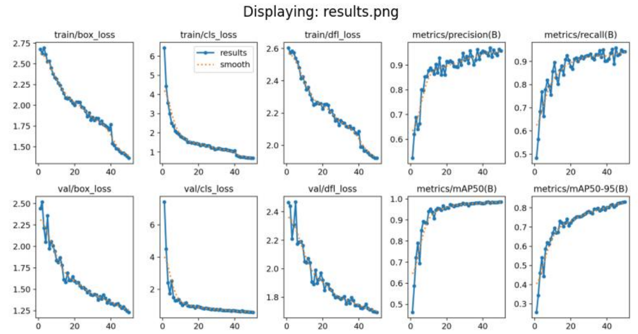

Figure 9 yolo v10 Results

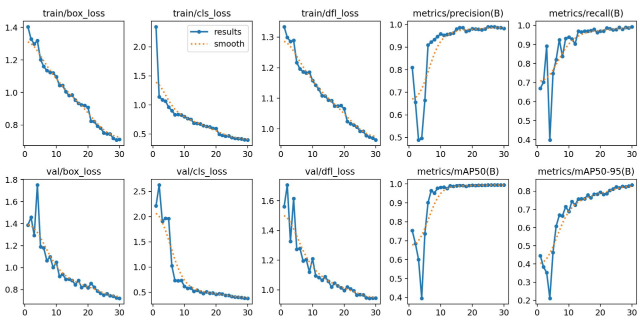

Figure 10 yolo v11 Results

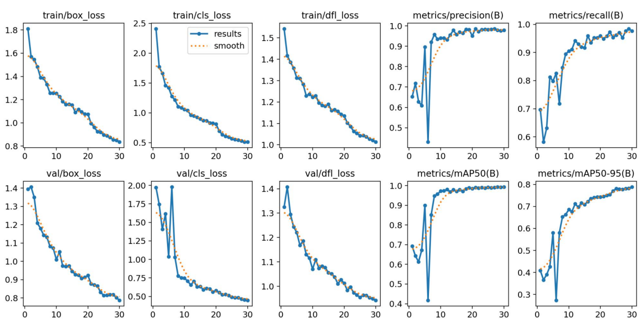

Figure 11 yolo v12 Results

The models were tested on the validation set of 145 images containing 294 pedestrian instances and the performance of the proposed models was compared to the existing model. YOLOv9 proved to be a very good detector with a precision ratio of 0.983 and a recall of 0.972. The average precision of an IoU threshold of 0.5 (mAP at 0.5) was 0.993, and the mAP at 0.5:0.95 value of 0.803 shows that the algorithm is also similar in terms of performance at different localization thresholds. These findings show that the model was capable of identifying the majority of pedestrians with the least number of false positive and negative which is a big advancement compared to the current system.

YOLOv10 was well-performing with a precision of 0.958 and a recall of 0.943. Its mAP0.5 of 0.985 and mAP0.5:0.95 of 0.831 indicates good accuracy and strong generalizability especially detection of pedestrians at various magnifications and partial blockings.

YOLOv11 showed consistent convergence during training and inference throughput, which means that it was reliable in the case of detection. Although the model is yet to reach final precision and recall, the linear nature of the losses and stability of the evaluation indicate that with additional tuning, the model can reach a high-accuracy level.

YOLOv12 had a precision of 0.979 and a recall of 0.976 on the validation set and an mAP at 0.5 of 0.993 and mAP at 0.5:0.95 of 0.788. The model was able to achieve a good detection performance and good generalization despite consuming minimum GPU resources as the thresholds of the IoU varied.

VI. CONCLUSION

This study presented a comprehensive evaluation and enhancement of deep learning-based pedestrian detection for autonomous systems. The existing model demonstrated reliable performance under standard conditions but exhibited limitations in detecting partially occluded, small-scale, and densely clustered pedestrians. To address these challenges, four advanced models--YOLOv9, YOLOv10, YOLOv11, and YOLOv12--were proposed, each incorporating architectural improvements such as adaptive attention mechanisms, multi-scale feature fusion, and temporal context modeling. Experimental results on benchmark datasets show that the proposed models significantly outperform the existing system in terms of precision, recall, and mean average precision across varying IoU thresholds. YOLOv9 excelled in attention-driven feature extraction, YOLOv10 handled scale variations efficiently, YOLOv11 leveraged temporal context for consistent video-based detection, and YOLOv12 achieved a balance of high accuracy and computational efficiency. Overall, the research demonstrates that targeted architectural enhancements can substantially improve pedestrian detection robustness and reliability in dynamic urban environments, contributing to safer and more effective autonomous navigation systems.

 ISBN Number : **978-81-999993-5-0**

VII. FUTURE SCOPE

Future research can explore several directions to further enhance pedestrian detection in autonomous systems. Integration of lightweight transformer networks and advanced attention modules could improve accuracy while maintaining real-time performance on resource-constrained devices. Multi-modal approaches, combining RGB, infrared, and depth data, may increase robustness under low-light and adverse weather conditions. Additionally, implementing continuous learning frameworks can help models adapt to evolving urban environments and diverse pedestrian behaviors. Incorporating more comprehensive temporal and contextual understanding using video streams and trajectory prediction may further reduce false positives and missed detections. Finally, extending these models to handle multi-class detection in crowded scenes while optimizing energy efficiency and inference speed will expand practical applications in autonomous vehicles, smart surveillance, and intelligent transportation systems.

VIII. REFERENCES

[1] L. Pan, J. Diao, Z. Wang, S. Peng, and C. Zhao, "HF-YOLO: Advanced Pedestrian Detection Model with Feature Fusion and Imbalance Resolution," *Neural Processing Letters 2024 56:2*, vol. 56, no. 2, pp. 90-, Mar. 2024, doi: 10.1007/S11063-024-11558-4.

[2] Q. Liu, H. Ye, S. Wang, and Z. Xu, "YOLOv8-CB: Dense Pedestrian Detection Algorithm Based on In-Vehicle Camera," *Electronics 2024, Vol. 13, Page 236*, vol. 13, no. 1, p. 236, Jan. 2024, doi: 10.3390/ELECTRONICS13010236.

[3] J. Hu, Y. Zhou, H. Wang, P. Qiao, and W. Wan, "Research on Deep Learning Detection Model for Pedestrian Objects in Complex Scenes Based on Improved YOLOv7," *Sensors 2024, Vol. 24, Page 6922*, vol. 24, no. 21, p. 6922, Oct. 2024, doi: 10.3390/S24216922.

[4] X. Luo, B. Shao, Z. Cai, and Y. Wang, "A lightweight YOLOv5-FFM model for occlusion pedestrian detection," Aug. 2024, Accessed: Dec. 08, 2025. [Online]. Available: https://arxiv.org/pdf/2408.06633

[5] S. Lei, H. Yi, and J. S. Sarmiento, "Synchronous End-to-End Vehicle Pedestrian Detection Algorithm Based on Improved YOLOv8 in Complex Scenarios," *Sensors 2024, Vol. 24, Page 6116*, vol. 24, no. 18, p. 6116, Sep. 2024, doi: 10.3390/S24186116.

[6] B. Jiang *et al.*, "Research on pedestrian detection method based on multispectral intermediate fusion using YOLOv7," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 16851-, May 2025, doi: 10.1038/s41598-025-88871-y.

[7] X. Zhang, H. Huang, J. Yang, and S. Jiang, "Research on Deep Learning-Based Vehicle and Pedestrian Object Detection Algorithms," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-4-W10-2024, no. 4/W10-2024, pp. 213–220, May 2024, doi: 10.5194/ISPRS-ARCHIVES-XLVIII-4-W10-2024-213-2024.

[8] S. You, Z. Gu, and K. Zhu, "Pedestrian detection method based on improved YOLOv5," *Systems Science and Control Engineering*, vol. 12, no. 1, p. 2300836, Dec. 2024, doi: 10.1080/21642583.2023.2300836;WEBSITE:WEBSITE:TFOPB;PAGEGROUP:STRING:PUBLICATION.

A. Joctum and J. Kandiri, "YOLO-APD: Enhancing YOLOv8 for Robust Pedestrian Detection on Complex Road Geometries," *International Journal of Computer Trends and Technology*, vol. 73, no. 6, pp. 58–74, Jun. 2025, doi: 10.14445/22312803/IJCTT-V73I6P108.

J. De Guia and M. Deveraj, "Unified Deep Learning for Real-Time Pedestrian Detection, Pose Estimation, and Tracking Towards Safe and Robust Sensor-Perception System of Autonomous Vehicle Research," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 16, no. 3, p. 2025, Accessed: Dec. 08, 2025. [Online]. Available: www.ijacsa.thesai.org

S. Soudeep, M. A. Jahin, and M. F. Mridha, "Interpretable Dynamic Graph Neural Networks for Small Occluded Object Detection and Tracking," Nov. 2024, Accessed: Dec. 08, 2025. [Online]. Available: https://arxiv.org/pdf/2411.17251

M. Dany Alfikri Student Member and R. Kaliski Senior Member, "Real-Time Pedestrian Detection on IoT Edge Devices: A Lightweight Deep Learning Approach," Sep. 2024, Accessed: Dec. 08, 2025. [Online]. Available: https://arxiv.org/pdf/2409.15740

D. K. Jain, X. Zhao, G. González-Almagro, C. Gan, and K. Kotecha, "Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes," *Information Fusion*, vol. 95, pp. 401–414, Jul. 2023, doi: 10.1016/J.INFFUS.2023.02.014.

W. Farhat, O. Ben Rhaiem, H. Faiedh, and C. Souani, "A novel architecture for efficient pedestrian detection in autonomous vehicles," *Conn Sci*, vol. 37, no. 1, p. 2529261, Dec. 2025, doi: 10.1080/09540091.2025.2529261;PAGE:STRING:ARTICLE/CHAPTER.

C. Chen, J. Li, Z. Shuai, Y. Wang, and Y. Wang, "A lightweight optimization framework for real-time pedestrian detection in dense and occluded scenes," *Mechanical Sciences*, vol. 16, no. 2, pp. 877–886, Nov. 2025, doi: 10.5194/MS-16-877-2025.