A stereo-vision system for real-time person detection in ADAS applications using a fine-tuned version of YOLOv5

Oumayma Rachidi, Chafik Ed-Dahmani, Badr Bououlid Idrissi
Department of Electromechanical Engineering, School of Arts and Crafts, Moulay Ismail University, Meknes, Morocco

Article Info

Article history:

Received Feb 29, 2024
Revised Aug 26, 2024
Accepted Sep 4, 2024

Keywords:

Advanced driver assistance systems
Computer vision
Deep learning
Person detection
Raspberry Pi
Stereovision
You only look once version 5

ABSTRACT

Pedestrian detection holds significant importance in advanced driver assistance systems (ADAS) applications, and presents a challenging task in this field. While the advent of deep learning has facilitated the introduction of various pedestrian detectors characterized by accuracy and low inference speed, there persists a need for further improvements. Notably, ADAS requires accurate detection of pedestrians in various environmental conditions that can adversely impact the model's performance, such as poor lighting, and bad weather. Furthermore, an imperative requirement involves the incorporation of distance estimation in conjunction with pedestrian detection, with an extension of detection capabilities to encompass cyclists and riders, who are equally crucial for ensuring road safety. Therefore, this paper introduces a stereovision system designed for the detection of pedestrians, cyclists, and riders. The initial phase, involves improving the performance of you only look once version 5 (YOLOv5s) through a fine-tuning process with a custom dataset integrating augmentation techniques to common objects in context (COCO) dataset. The detector is trained using Google Colab, and tested in real-time with a Raspberry Pi 4 model B, 8 G RAM. A comparative analysis is conducted between the YOLOv5s and the fine-tuned model to prove the accuracy of our approach. The results showcase a high performance of the detector reaching an accuracy exceeding 79%.

Corresponding Author:

Oumayma Rachidi
Department of Electromechanical Engineering, School of Arts and Crafts, Moulay Ismail University
Marjane 2, BP 15290, Al-Mansour, Meknes, Morocco
Email: oum.rachidi@edu.umi.ac.ma

1. INTRODUCTION

Object detection constitutes a fundamental component within the field of computer vision, playing a crucial role in various applications, such as surveillance, medical and health care systems, and advanced driver assistance systems (ADAS). In the context of ADAS, the precise identification and localization of obstacles is imperative to avoid collisions and ensure a safe driving. Among the paramount of obstacles in this context are persons, encompassing pedestrians, cyclists, and riders, necessitating accurate and fast detection to facilitate reliable decision-making processes.

The pursuit of reliable detection mechanisms has posed challenges over the last decades, prompting the design of various systems equipped with different sensors such as LIDAR, RADAR, and cameras [1]. Notably, recent advancements have focused on vision-only systems, leveraging monocular or stereo-cameras. Stereovision, in particular, has gained prominence due to its cost effectiveness and capacity to emulate human-like patterns for visual perception. A salient aspect of this technique is its capacity for depth estimation, providing critical information regarding the object's distance in relation to the vehicle. The integration of object detection and distance estimation is crucial in ADAS applications, as it equips the system with comprehensive data for executing necessary actions and ensuring safe driving practices.

Historically, the identification of pedestrians necessitated a manual extraction of features through a sliding window approach, followed by feeding these features into a classifier. Among the traditional hand-designed features are mainly Haar, Harris and Stephens [2], speeded up robust features (SURF) [3], Haar wavelet [4], Hu moment [5], histogram of oriented gradients (HOG) [6]. Moreover, classification methodologies are typically categorized into supervised and unsupervised algorithms. Supervised algorithms commonly employ the support vector machine (SVM) [7] and Naive Bayes classifier or perception [8]. In contrast, unsupervised algorithms typically utilize mean shift [9] and K-means [10]. Traditional approaches for pedestrian detection are limited in complex scenes, as they necessitate manual setup and region selection through the sliding window technique.

The advent of deep learning has seen the introduction of numerous models for object detection, which can be categorized into two main types: two-stage and one-stage detectors. In the primary classification, regions of interest are initially generated, and subsequently input into a network for classification and bounding boxes regression. While these detectors demonstrate high accuracy, their processing time poses challenges for real time applications. Prominent examples in this category include region-based convolutional neural network (R-CNN) [11], fast R-CNN [12], and faster R-CNN [13]. Conversely, one-stage detectors perform object classification and bounding boxes regression through a single network, sacrifying some accuracy for enhanced inference speed. Exemplary models in this category, include the single shot multi box detector (SSD) [14], and the you only look once (YOLO) [15]. Since its inception, YOLO has gained considerable attention and continued to evolve across different versions. Notably, YOLOv5, representing one of the recent versions, has demonstrated its effectiveness in real-time applications.

Pedestrian detection emerges as a critical task within the domain of computer vision, presenting inherent challenges. Recently, many research have addressed one or some of the challenging factors such as occlusion, low-quality image, and detection under various lighting conditions. The majority of these research have been focusing on the utilization of deep neural networks. Zhang *et al.* [16] designed an occlusion-aware R-CNN to enhance the detector's accuracy within crowded scenes. This model incorporated an aggregation loss and a dedicated region of interest (ROI) pooling layer to effectively detect partially occluded pedestrians. Furthermore, within the domain of partial occlusion, the DeepParts model was introduced, incorporating forty-five distinct part detectors capable of discerning partial occluded pedestrian. The issue of occlusion challenge is frequently examined individually; hence a joint deep learning framework was developed to concurrently acquire proficiency in occlusion handling, along with tasks such as feature extraction, classification and deformation handling [17]. Hou *et al.* [18] developed multi spectral pedestrian detector based on SSD, prompting the utilization of thermal images to address pedestrian detection challenges during nighttime scenarios. In the same context, Li *et al.* [19] proposed a hybrid pedestrian detection approach integrating both RGB and thermal images. The model employed two backbone networks based on YOLOv5 and integrated them through a fusion feature selection module to generate a fusion map. Exhibiting superior performance in comparison to advanced methods, the model has demonstrated its efficiency in detecting pedestrians under challenging lighting conditions. In addressing the challenge of detecting small objects within infrared images, a model based on YOLOv5 was proposed [20]. Significant modifications to the original model architecture, loss function, and training strategy yielded favorable results in both human and vehicle detection. Notably, Al-Tameemi *et al.* [21] have recently used the YOLOv5 pretrained model in a monitoring robotic system, demonstrating real-time object recognition capabilities. Recently, Khan *et al.* [22] introduced the fast focal detection network (F2DNET), specifically designed to enhance the performance of two stage detectors. The proposed network redesigns the two-stage detection architecture, involving the replacement of the region proposal network with a robust focal detection network, complemented by a fast suppression head to mitigate false positives. Operating as an anchor free detection network, it relies on center and scale map prediction. Comparative assessments against state-of-the-art demonstrate its efficacy across widely utilized datasets: Euro city persons, Caltech, and city persons datasets. Additionally, Khan *et al.* [23] introduced the localized semantic feature mixers for efficient pedestrian detection in autonomous driving (LSFM). Designed to address challenges associated with small or heavily occluded pedestrians, this model incorporates a ConvMLP-based backbone, a super pixel pyramid pooling neck for feature filtering and enrichment, and a dense focal detection network. LSFM outperforms all recent detectors, including F2DNet across: Euro city persons, Caltech, and city persons datasets. However, it is important to note that the LSFM model was exclusively evaluated in daytime scenes and traffic scenarios. While these models exhibit promising advancements, their efficacy in real-time applications is yet to be validated due to their recent introduction.

In contrast to the mentioned studies, our work seeks to fine-tune the YOLOv5 architecture specifically for person detection, prioritizing simplicity and efficiency. The primary objective is to develop an accurate detector capable of identifying persons in various environmental conditions. While a significant portion of ADAS research concentrates on the detection of pedestrians, it is imperative to extend this focus to include cyclists and riders for comprehensive road safety. Furthermore, limited attention has been devoted to the integration of distance calculation into object detection methodologies, particularly with the application of stereovision techniques. Therefore, this paper aims to present a stereovision system specifically designed to detect persons in roadways. The main contributions of our study are: i) design of a stereovision system using an economical, energy-efficient, and highly effective device, ii) an approach for real-time person detection across various contexts using transfer learning, embedded on the same economical, energy-efficient, and highly effective artificially intelligent (AI) device, and iii) a custom dataset created to train the model using augmentation techniques to develop a robust model for person detection.

The paper is organized as follows: section 2 outlines the research method, emphasizing camera calibration techniques and the application of transfer learning for the development of the new detector. Section 3 comprehensively discusses the results and experiments findings. Finally, section 4 presents the principal outcomes and offers suggestions for future research directions.

2. METHOD

The devised system was designed in our laboratory utilizing a 3D printer. The stereoscopic camera, consisting of two OV5647 5MP cameras, was developed as a component of this system. Operational control is facilitated by a Raspberry Pi 4 Model B, employing computer vision techniques to provide real-time video incorporating person detection based on the fine-tuned YOLOv5s model. The comprehensive design is visually depicted in Figure 1. This research centers on the stereo camera calibration and the transfer learning technique for person detection, specifically designed for application in ADAS.

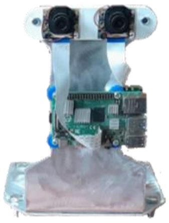


Figure 1. The proposed system

2.1. Hardware components

2.1.1. Raspberry Pi 4 model B

A recently introduced Raspberry Pi kit, designed for invoking contemporary AI models with considerations for small-scale deployment, low power consumption, enhanced speed, and cost-effectiveness, has been employed in this study [24]. This version, denoted as the Raspberry Pi 4 model B with 8 GB of memory, as depicted in Figure 2, incorporates general purpose input-output (GPIO) pins, a camera serial interface (CSI) port, and two micro-HDMI terminals. Operated by a type C mini-USB, this microcomputer board has become a platform conductive to the realization of various new detectors accommodating diverse AI workloads. Its compatibility with 5V power supply renders it an energy-efficient and low-power embedded device. Additionally, a 128 GB secure digital (SD) memory card is utilized for operating system storage and enabling the handling of large volumes of data for both reading and writing.

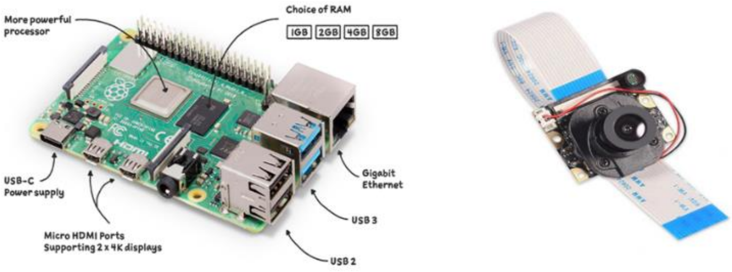


Figure 2. Hardware components

2.1.2. OV5647 5MP cameras

Manufactured by Omnivision Technologies, the OV5647 camera module provides a resolution of 2592×1944 pixels, resulting in the acquisition of images characterized by a high degree of sharpness and detail. Employing a complementary metal-oxide-semiconductor (CMOS) image sensor, the camera offers a low power consumption and high-quality imaging capabilities. Its integration with the Raspberry Pi is facilitated through the CSI port. Following the successful interfacing of the camera with the Raspberry Pi, a calibration procedure is executed. In the context of this research, a stereovision system was designed utilizing two OV5647 5MP cameras. Each camera is individually interfaced with a Raspberry Pi, necessitating the implementation of a sockets transmission to consolidate the computational framework into a singular Raspberry Pi unit.

2.2. Stereo camera calibration

Stereo cameras provide unique benefits in comparison to other sensors frequently utilized in ADAS systems, including RADAR, LIDAR, and monocular cameras [25]. Specifically, their ability to acquire high-resolution images containing detailed visual information, including colors and shapes, proves essential for applications related to object detection. Stereovision is particularly advantageous in situations characterized by occlusion, and it facilitates the creation of a three-dimensional reconstruction of the scene, thereby supporting precise distance estimation.

Within the domain of computer vision, camera calibration stands out as a crucial task. This procedure results in the derivation of intrinsic and extrinsic camera parameters. Intrinsic parameters serve to map camera coordinates into the image plane, supplying vital details for comprehending the scene's geometry. Simultaneously, extrinsic parameters convey the camera's orientation with respect to the world coordinate frame. In stereovision applications, the accurate determination of depth and distance heavily relies on these parameters. Calibration of the two cameras becomes imperative to achieve precise distance estimation. Throughout the calibration procedure, a recognized geometric object is employed, and equations derived from the object's coordinates in the world reference and on the image plane aid in establishing the camera's parameters. In our investigation, we utilized OpenCV for the stereo camera calibration process. As depicted in Figure 3, chessboard images were concurrently generated from both the left and right cameras. The intrinsic and extrinsic parameters were then obtained by applying predefined functions available within the OpenCV library. Subsequently, the acquired parameters were incorporated into image processing techniques to eliminate distortion and rectify the captured images.

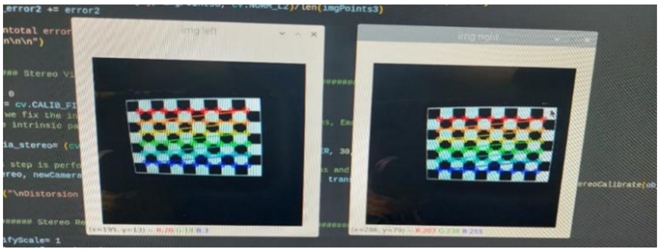


Figure 3. Left and right images taken during the stereo camera calibration process using OpenCV

2.3. You only look once version 5 architecture

YOLOv5 object detection model proposed in this work, represents a noteworthy iteration in the evolution of object detection methodologies within the field of computer vision. It has gained substantial attention for its innovative architectural design, which combines efficiency, accuracy, and real-time processing capabilities [26]. The model incorporates a CSPDarknet53 as a backbone, contributing to improved feature extraction capabilities, and a path aggregation network (PANet) for enhancing feature fusion across different scales, enabling robust object detection, and allowing YOLOv5 to excel in scenarios with varying object scales and complexities. Figure 4 [27] presents the model architecture: the fundamental convolution module is presented by the CBL, and the Bottleneck cross-stage partial (CSP) is used to efficiently aggregate and fuse features across stages, enhancing the model's capacity for capturing intricate spatial hierarchies and facilitating robust object detection in diverse scenarios. The selection of YOLOv5s for person detection in this work, is based on its superior trade-off between accuracy and computational efficiency, which is essential for resource-constrained environments. YOLOv5s offers a lightweight model with faster inference times and lower computational demands, making it ideal for devices with limited processing power, such as the Raspberry Pi. Compared to other detectors, such as faster R-CNN and mask

R-CNN [28], which provide high accuracy but require substantial computational resources, YOLOv5s is more suitable for real-time applications. Other examples include SSD, which offers a compromise between speed and accuracy but does not match the precision and efficiency of YOLOv5s, and RetinaNet [29], which, although effective, can be computationally intensive. Additionally, YOLOv5s benefits from a streamlined architecture with advanced features like CSPNet and PANet, enhancing feature extraction and information flow. This ensures robust detection performance with minimal latency. In contrast, its larger counterparts, YOLOv5m [30], YOLOv5l [31], and YOLOv5x [32], while providing higher accuracy, are more computationally demanding. These characteristics make YOLOv5s the optimal choice for this study.

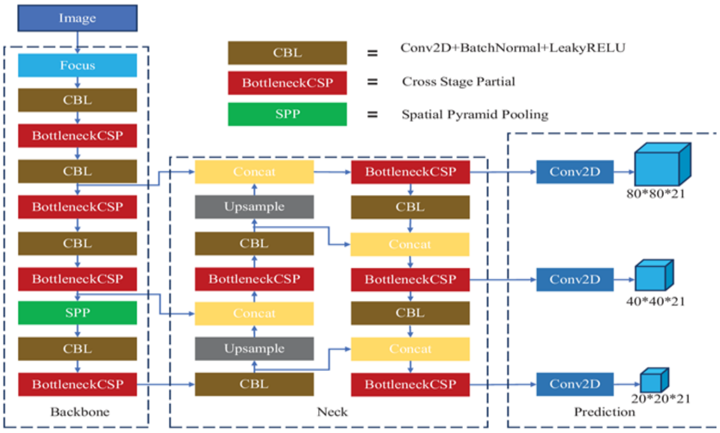


Figure 4. YOLOv5 model architecture

2.4. Dataset

A custom dataset for person detection was generated by building upon the common objects in context (COCO) dataset [33]. Approximately, 40,000 training images and 2,758 validation images were selected from the COCO dataset, with a specific emphasis on the person class. To enhance the robustness of the detector and address common challenges in person detection, extensive data augmentation techniques were employed. Specifically, Roboflow was utilized for the construction and augmentation of our dataset. The applied augmentations to the initial images include:

- Brightness: a 20% decrease in brightness was implemented as an augmentation technique to improve the model’s resilience to lighting challenges, particularly during nighttime scenes.
- Blur: a Gaussian blur of 2.5 pixel was applied to improve the model’s adaptability to camera focus issues. In the context of ADAS, where cameras are subject to movement, and pedestrians, cyclists and riders may be either in motion or stationary, this technique contributes to the model’s efficiency.
- Noise: a noise level up to 2% of pixels was applied to foster the model’s resilience to camera artifacts. This approach aids the model in assimilating various aspects of the person class by intentionally occluding random features.

By implementing these techniques, we generated a dataset of 118,425 training images, 2,758 validation images, and 186 testing images. Our dataset is available on Roboflow.

2.5. Model development

In this study, YOLOv5s was fine-tuned using the custom dataset to optimize both computational efficiency and model performance. This choice was primarily motivated by the advantage of leveraging pre-existing feature representations derived from extensive datasets, which significantly reduces the need for extensive computational resources and data compared to training from scratch. Fine-tuning enables accelerated convergence by utilizing robust initial weights and learned knowledge from comprehensive pre-training on datasets such as COCO. This approach allows for the effective harnessing of advanced model capabilities while maintaining computational feasibility, thereby achieving a balanced trade-off between performance and resource efficiency.

To implement and evaluate the fine-tuned model, the training process was conducted using Google Colab with Tesla processors to minimize training time. The model is stored for subsequent testing and validation on a Raspberry Pi 4 model B, equipped with 8 GB RAM. The proposed model for person detection from image or video frames is implemented using the Python PyTorch package.

To assess the efficacy of our methodology, we used the following metrics:

- Precision: denotes the proportion of correctly classified samples relative to the total number of predicted positive samples, encompassing both correct and incorrect classifications. This metric illustrates the accuracy of the model to classify a sample as positive, as mentioned in (1), where TP is true positives and FP is false positives.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

- Recall: denotes the proportion of correctly classified samples, to the total number of positive samples. This metric serves as a measure of the model's effectiveness in identifying positive samples, as mentioned in (2), where FN is false negatives.

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

- Average precision (AP): represents the area under the precision-recall curve, generated according to different confidence thresholds:

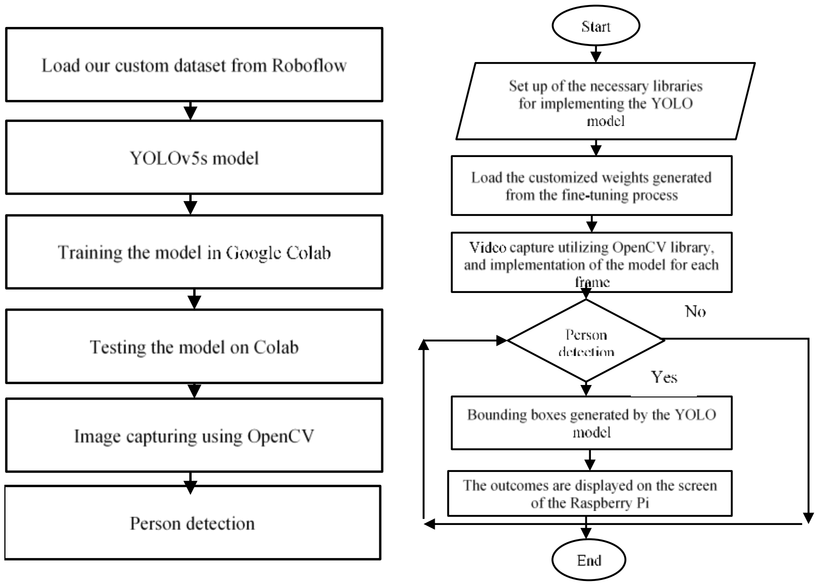
$$AP = \sum_{k=0}^{k=n-1} \left(R_C(k) - R_C(k+1)\right) x \, P_R(k) \tag{3}$$

where n is the number of thresholds, and P_R and R_C refer to precision and recall respectively.

- Mean average precision (mAP): denotes the mean of all AP through different object classes. In the context of our study, focusing exclusively on the person class only, this metric aligns with the AP.

2.6. The final setup

Following dataset creation and loading, the YOLOv5s model is trained using Google Colab GPU, resulting in a high accuracy, and a robust model able to detect persons in diverse contexts. The procedural steps of the training/testing phase are presented on Figure 5(a). Afterwards, the proposed system is used to implement the model in the Raspberry Pi and start real-time image/video capturing. To successfully execute the model, it is imperative to install deep learning libraries and frameworks on the hardware. Python served as the programming language for importing the essential packages. Real time video capture is facilitated by the OpenCV library, leveraging a connected camera with the Raspberry Pi. The camera is activated to capture frames, an algorithm is employed to transmit these frames to the model for person detection. A continuous loop is utilized to display the frames in real-time, incorporating the outcomes of person detection, until the user initiates an exit command by pressing the "q" key on the keyboard. Consequently, the real-time results of person detection are displayed on the Raspberry Pi. Figure 5(b) illustrates the key steps used in implementing the deep neural network in the Raspberry Pi.


Load our custom dataset from Roboflow
YOLOv5s model
Training the model in Google Colab
Testing the model on Colab
Image capturing using OpenCV
Person detection
Start
Set up of the necessary libraries for implementing the YOLO model
Load the customized weights generated from the fine-tuning process
Video capture utilizing OpenCV library, and implementation of the model for each frame
Person detection
No
Yes
Bounding boxes generated by the YOLO model
The outcomes are displayed on the screen of the Raspberry Pi
End

(a) (b)

Figure 5. Process flow of the proposed system; (a) training and testing and (b) implementation phase

3. RESULTS AND DISCUSSION

The evaluation of the model was executed through the utilization of Google Colab: the validation set was employed for the computation of validation metrics, and the results revealed the high performance of our fine-tuned model in detecting persons across diverse scenarios.

3.1. Training process

The model's training process was executed on the Google Colab platform, leveraging its robust GPU capabilities without requiring additional configuration. The YOLOv5s architecture was chosen for this training due to its balance of performance and efficiency. The training extended over 80 epochs, providing sufficient iterations for the model to effectively learn and generalize from the data. The training parameters were meticulously chosen to optimize the learning process:

- Learning rate (0.01): the learning rate determines the magnitude of updates applied to the model's weights during training. A learning rate of 0.01 was selected to achieve a suitable balance between convergence speed and stability, ensuring that the model makes significant progress without diverging from the optimal solution or converging too slowly.
- Momentum (0.937): momentum is a technique used to accelerate the convergence of the optimization process by incorporating information from previous gradient updates into the current weight adjustments. By setting the momentum to 0.937, the optimization process gains substantial inertia, which contributes to a smoother trajectory and mitigates oscillations in the weight updates. This strategy improves the model's capability to navigate through local minima, promoting accelerated and more consistent convergence by effectively utilizing accumulated gradient information to optimize weight updates.
- Batch size (16): the batch size refers to the number of training examples processed before updating the model's weights. Smaller batch sizes generally offer more frequent updates to the model weights, which can lead to better generalization, but may increase training time. Conversely, larger batch sizes can speed up training but might require more memory and may lead to less frequent updates. In this context, a batch size of 16 is chosen to optimize memory usage while maintaining a reasonable update frequency and training stability.

Roboflow was utilized to preprocess the dataset, converting it into the YOLOv5 PyTorch format and exporting it as a ZIP file. The images in the dataset were of 640×640 resolution, providing an appropriate level of detail for the person detection task while keeping computational demands manageable. The final model architecture comprises 7,012,822 parameters distributed across 157 layers, reflecting a well-balanced design aimed at achieving optimal performance while maintaining computational efficiency.

During the training process, several challenges emerged: ensuring effective convergence involved addressing inherent risks of overfitting, particularly with the extensive data augmentation applied. Different numbers of epochs were tested to identify the optimal duration for training, balancing between sufficient learning and avoiding overfitting. Computational constraints were another significant challenge, as even with Google Colab's Tesla GPUs, managing memory and processing power was critical to avoid out-of-memory errors and excessive training times. Additionally, the complexities introduced by the data augmentation techniques required careful management to maintain training stability and ensure that the model could effectively generalize. These challenges were systematically addressed to optimize the performance and reliability of the YOLOv5s model.

To assess the outcomes of the training process, Figure 6 displays ten curves that illustrate validation metrics across various training epochs. The top five curves represent the model's performance during training, whereas the lower set refer to the model's validation phase. Among these curves, the AP and mAP metrics are notable. The AP involves calculating precision over different intersection over union (IOU) thresholds, while the mAP represents the average AP across all classes. In this work, mAP is equivalent to AP. The mAP [.5:.95] metric signifies the average AP calculated across IOU thresholds ranging from 50% to 95% with an increment of 5%. Additionally, the figure includes the loss function, which is composed of three elements: Object loss, evaluating the algorithm's ability to predict the presence of an object; box loss, assessing the alignment of the bounding box with the ground truth, and class loss, reflecting the accuracy of the algorithm in predicting the object class. In the context of this study, the class loss consistently maintains a zero-value due to the singular class context. From the analysis of the figure, it is evident that all metrics demonstrate consistent improvement from the initial stages of the training process to its conclusion. This trend indicates the effective fine-tuning of the model, with the initial epochs reflecting the performance of the pretrained weights and the subsequent epochs illustrating progressive enhancements achieved through fine-tuning.

To further evaluate the detector's performance, the precision-recall curve was generated from different confidence score values as shown in Figure 7. A decline in the confidence score is linked to a proportional rise in the recall. Conversely, an elevation in the score is accompanied by an increase in precision. Thus, there is a balance between precision and recall that depends on the chosen confidence threshold. The precision-recall curve indicates that the YOLOv5 model achieves high precision and recall rates in the task of person detection.

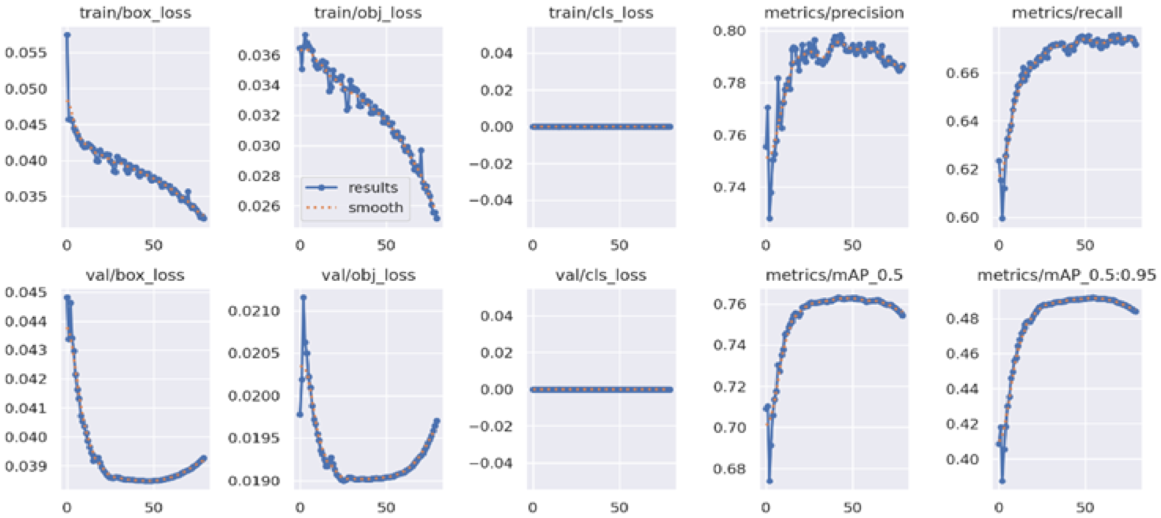


Figure 6. Fluctuations of key metrics throughout different training epochs

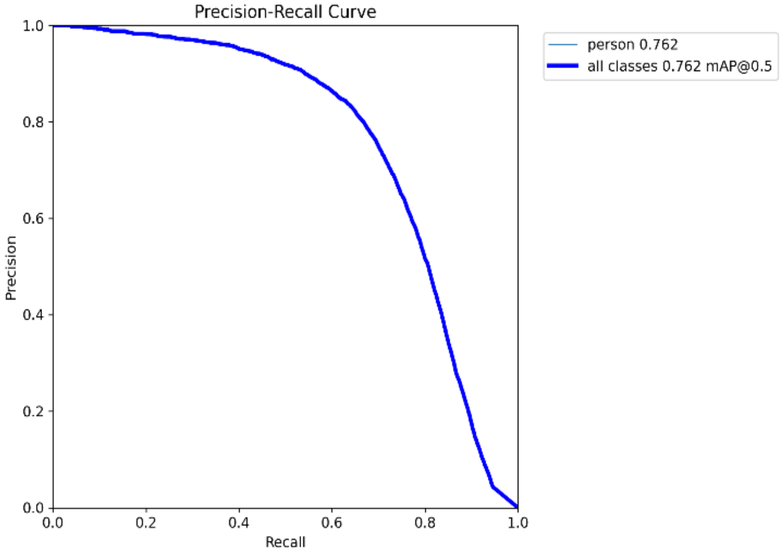


Figure 7. The precision recall curve of our approach

Table 1 displays the outcomes derived from our proposed method for person detection. Our model achieves a precision of 79.1% and a recall of 67.3%. The mAP at 0.5 IOU threshold stands at approximately 76%, while the mAP [.5:.95] reaches around 49%. To further underscore the effectiveness of our approach, we conducted a visual comparison between our method and the pretrained YOLOv5s model on the COCO dataset. Figure 8 presents the comparative results, clearly demonstrating the enhanced accuracy in person detection achieved through our fine-tuning process.

Table 1. Experimental results of the fine-tuned model

Model	Precision	Recall	mAP 0.5	mAP [.5:.95]
YOLOv5s (proposed)	0.791	0.673	0.762	0.492

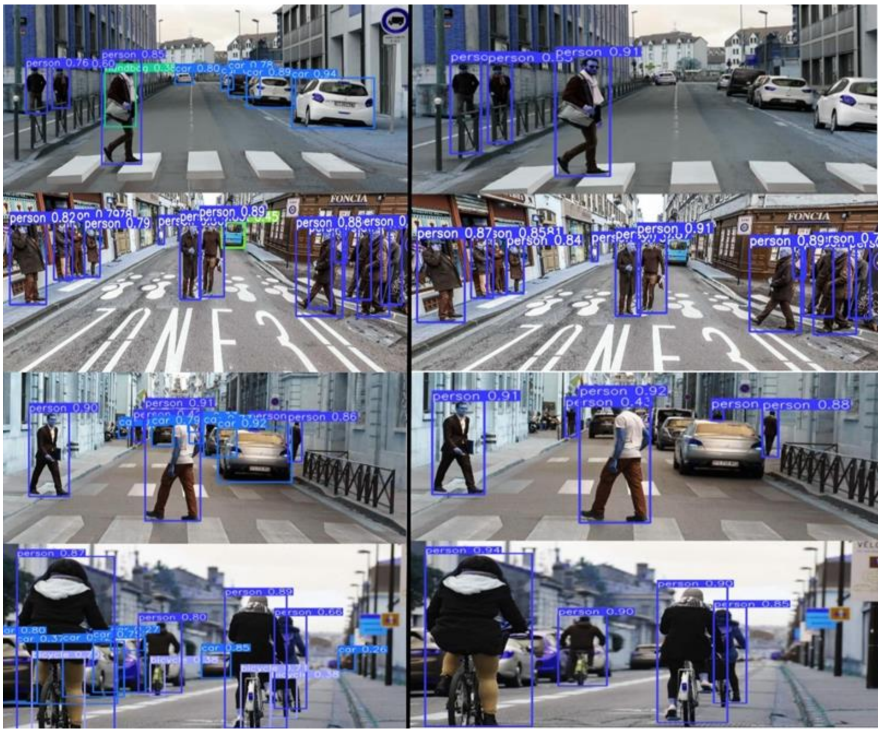


Figure 8. Visual comparison between pretrained YOLOv5s and our approach: the detections on the left are generated with YOLOv5s pretrained model and the detections on the right are produced by our approach

To further enhance the results, several techniques could be considered for future iterations of the training process, such as dropout and adaptive learning rates. Dropout can improve model performance by reducing overfitting, as it prevents the model from becoming overly reliant on specific features, thereby promoting better generalization to unseen data. Additionally, employing adaptive learning rate methods, such as Adam, could optimize training outcomes by adjusting the learning rate based on gradient statistics. This approach facilitates more efficient and stable convergence, which may lead to more accurate predictions. Implementing these strategies could refine the model's training process and enhance its overall effectiveness.

3.2. Model implementation

Following the training and testing phases, the model is deployed and validated on the designated embedded device utilizing the camera Pi module. The efficacy of our novel approach was evaluated in real-time using our system, highlighting the model’s capability for real-time person detection in diverse scenarios. Figure 9 showcases selected frames captured from the camera, illustrating the robust performance of the detector. Therefore, this paper elucidates the impact of the fine-tuning process on the performance of the detector. The incorporation of pretrained weights with our dataset, derived from the COCO dataset with augmentation techniques, enhances the model’s performance in detecting persons across various environmental conditions. Additionally, we conducted real-time image testing to validate the detector’s efficiency, affirming its robustness and suitability for real-time applications.

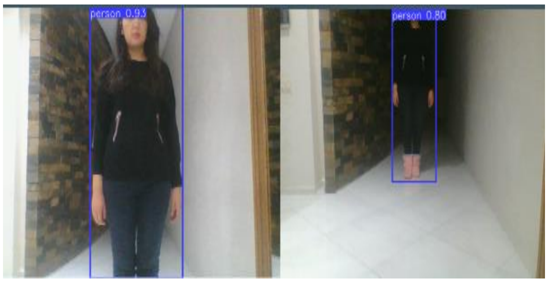


Figure 9. Selected frames from real time person detection using the fine-tuned model

4. CONCLUSION

Ensuring a safe distance between vehicles and road users is one of the most important tasks in ADAS applications. The detection of pedestrians, cyclists, and riders poses considerable challenges within this domain, involving both person detection and distance estimation. This paper presented a stereovision system specifically designed for person detection within the context of ADAS. YOLOv5s was selected as a detector due to its favorable trade-off between accuracy and speed, with a subsequent fine-tuning process implemented to enhance the detector's performance. The dataset utilized for this purpose derived from the COCO dataset for person class, including augmentation techniques to enhance the detector's accuracy in various environmental conditions. The detector was trained on Google Colab, and deployed on a Raspberry Pi 4 to perform person detection with real-time video captures. Consequently, this study validates the possibility of deploying the detector within highly constrained computational resources.

The obtained results reveal a high accuracy, with a mAP 0.5 reaching 76.2% and a Map [.5:.95] of 49,2%. Further work, will extend to encompass depth estimation utilizing the stereo-camera, as well as exploring the utilization of a higher computational resource instead of the Raspberry Pi. This exploration aims to further enhance the processing speed of the detector, making it well-suited for real-time applications. Additionally, adjustments will be made to implement the system in an electrical vehicle, simulating diverse functionalities along with person detection and distance estimation. The system will prioritize the detection of critical instances for reliable collision avoidance.

ACKNOWLEDGEMENTS

The authors wish to thank the University of Moulay Ismail for supporting this research.

REFERENCES

[1] M. Shirpour, N. Khairdoost, M. A. Bauer, and S. S. Beauchemin, "Traffic Object Detection and Recognition Based on the Attentional Visual Field of Drivers," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 594–604, 2023, doi: 10.1109/TIV.2021.3133849.

[2] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," in *British Machine Vision Association and Society for Pattern Recognition*, Alvey Vision Club, 1988, p. 147-151, doi: 10.5244/c.2.23.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008, doi: 10.1016/j.cviu.2007.09.014.

[4] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, IEEE, Jun. 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587800.

[5] M. K. Hu, "Visual Pattern Recognition by Moment Invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962, doi: 10.1109/TIT.1962.1057692.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, IEEE, 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[7] J. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," *Advances in Kernel Methods-Support Vector Learning*, 1998.

[8] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, doi: 10.1037/h0042519.

[9] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002, doi: 10.1109/34.1000236.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281–297.

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

[12] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

[13] 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, 2017, doi: 10.1109/TPAMI.2016.2577031.

[14] 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)*, vol. 9905 LNCS, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

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

[16] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-Aware R-CNN: Detecting Pedestrians in a Crowd," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11207 LNCS, 2018, pp. 657–674, doi: 10.1007/978-3-030-01219-9_39.

[17] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, Dec. 2015, pp. 1904–1912, doi: 10.1109/ICCV.2015.221.

[18] Y. L. Hou, Y. Song, X. Hao, Y. Shen, and M. Qian, "Multispectral pedestrian detection based on deep convolutional neural networks," in *2017 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2017*, IEEE, Oct. 2017, pp. 1–4, doi: 10.1109/ICSPCC.2017.8242507.

[19] M. Li, B. Liu, J. Sun, G. Zhang, and W. Su, "Multimodality pedestrian detection based on YOLOv5," in *International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP 2022)*, P. Loskot, Ed., SPIE, Nov. 2022, p. 51, doi: 10.1117/12.2659653.

[20] J. Kim, J. Huh, I. Park, J. Bak, D. Kim, and S. Lee, "Small Object Detection in Infrared Images: Learning from Imbalanced Cross-Domain

Data via Domain Adaptation," *Applied Sciences (Switzerland)*, vol. 12, no. 21, p. 11201, Nov. 2022, doi: 10.3390/app122111201.

[21] M. I. Al-Tameemi, A. A. Hasan, and B. K. Oleiwi, "Design and implementation monitoring robotic system based on you only look once model using deep learning technique," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 106–113, Mar. 2023, doi: 10.11591/ijai.v12.i1.pp106-113.

[22] A. H. Khan, M. Munir, L. V. Elst, and A. Dengel, "F2DNet: Fast Focal Detection Network for Pedestrian Detection," in *Proceedings - International Conference on Pattern Recognition*, IEEE, Aug. 2022, pp. 4658–4664, doi: 10.1109/ICPR56361.2022.9956732.

[23] A. H. Khan, M. S. Nawaz, and A. Dengel, "Localized Semantic Feature Mixers for Efficient Pedestrian Detection in Autonomous Driving," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2023, pp. 5476–5485, doi: 10.1109/CVPR52729.2023.00530.

[24] F. Mendoza-Cardenas, R. S. Leon-Aguilar, and J. L. Quiroz-Arroyo, "CP-ABE encryption over MQTT for an IoT system with Raspberry Pi," in *2022 56th Annual Conference on Information Sciences and Systems, CISS 2022*, IEEE, Mar. 2022, pp. 236–239, doi: 10.1109/CISS53076.2022.9751194.

[25] D. Ridel, P. Shinzato, A. R. Pereira, V. Grassi, and D. Wolf, "Obstacle avoidance using stereo-based generic obstacle tracking," in *Proceedings - 2017 LARS 14th Latin American Robotics Symposium and 2017 5th SBR Brazilian Symposium on Robotics, LARS-SBR 2017 - Part of the Robotics Conference 2017*, IEEE, Nov. 2017, pp. 1–6, doi: 10.1109/SBR-LARS-R.2017.8215284.

[26] O. Rachidi, E. D. Chafik, and B. Bououlid, "Design of a real-time-integrated system based on stereovision and YOLOv5 to detect objects," in *Enhancing Performance, Efficiency, and Security Through Complex Systems Control*, 2024, pp. 283–297, doi: 10.4018/979-8-3693-0497-6.ch016.

[27] F. Zhou, H. Zhao, and Z. Nie, "Safety Helmet Detection Based on YOLOv5," in *Proceedings of 2021 IEEE International Conference on Power Electronics, Computer Applications, ICPECA 2021*, IEEE, Jan. 2021, pp. 6–11, doi: 10.1109/ICPECA51329.2021.9362711.

[28] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[29] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[30] B. Liu, H. Wang, Y. Wang, C. Zhou, and L. Cai, "Lane Line Type Recognition Based on Improved YOLOv5," *Applied Sciences (Switzerland)*, vol. 13, no. 18, p. 10537, Sep. 2023, doi: 10.3390/app131810537.

[31] J. Li *et al.*, "A novel small object detection algorithm for UAVs based on YOLOv5," *Physica Scripta*, vol. 99, no. 3, p. 036001, Mar. 2024, doi: 10.1088/1402-4896/ad2147.

[32] K. V. Houde, P. M. Kamble, and R. S. Hegadi, "Trees Detection from Aerial Images Using the YOLOv5 Family," *Communications in Computer and Information Science*, pp. 314–323, 2024, doi: 10.1007/978-3-031-53082-1_25.

[33] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3686–3693.

BIOGRAPHIES OF AUTHORS

Oumayma Rachidi was born in Rabat, Morocco. She is a third year Ph.D. student and received an Engineer's degree from National Graduate School of Arts and Crafts, Meknes, Morocco, in 2021. She is actually working as an electromechanical engineer in the industry field. Her ongoing dissertation studies the pedestrian detection in ADAS systems, and explores deep learning techniques for object detection. She is particularly interested in industrial control systems, FPGA based ADAS systems, and electrical vehicles. She can be contacted at email: oum.rachidi@edu.umi.ac.ma.

Chafik Ed-Dahmani received his engineering in Mechatronics from the University of Abdelmalek Essaadi in 2014 and his Ph.D. degree from the University of Mohamed 5-Rabat, Morocco. Currently, he is an assistant professor at the National Graduate School of Arts and Crafts-Meknès, Morocco. His research interests include renewable energy system conversion, control systems, and microgrids. He can be contacted at email: c.eddahmani@umi.ac.ma.

Badr Bououlid Idrissi was born in Marrakech, Morocco. He received the Ph.D. degree from Faculté Polytechnique de Mons, Mons, Belgium, in 1997 and the Engineer's degree from Ecole Nationale de l'Industrie Minérale, Rabat, Morocco, in 1992. Since 1999, he has been working at Ecole Nationale Supérieure d'Arts et Métiers (ENSAM-Meknès), Moulay Ismaïl University, Meknès, Morocco, where he is a Professor in the Department of Electromechanical Engineering, in the areas of power electronics and electrical machines. His research interests are mainly electric drives, industrial control systems, DSP/FPGA based ADAS systems, and electrical vehicles. He can be contacted at email: b.bououlid@umi.ac.ma.