XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

Computer Vision Applied to the Management and Optimization of Logistics Operations

Marcos Antonio Nunes da Silveira[1], Gustavo Maia de Almeida[2]

[1]*Federal Institute of Espírito Santo, Serra*
marcosilveiraa@gmail.com
[2]*Federal Institute of Espírito Santo, Serra*
gmaia@ifes.edu.br

Abstract. This paper proposes a computer vision system for active monitoring of logistics operations, focusing on detecting and mitigating risks, such as interactions between people and vehicles. The main objective is to increase safety in the operational environment. IP cameras with RTSP transmission are used. Images are processed remotely on a notebook connected to the Wi-Fi network, while visual alerts are generated locally on the identified target. The use of color-coded segmentation in HSV space allows dynamically defining regions of interest (ROIs), with real-time adjustments using trackbars. Segmented objects have their contours detected and receive blue ROIs with configurable offsets. Target classification as ”person” or ”vehicle” is performed by a Convolutional Neural Network (CNN) with the YOLOv8n model. The preprocessing of the image dataset used the transfer learning technique via Roboflow, using 1,500 images. Data augmentation was applied, generating a final dataset of 3,612 images. The system was implemented in Python, using OpenCV and integration with ESP32 via HTML requests, triggering an alert LED. The combination of color segmentation in HSV space and CNN offers an accurate, efficient, and low-cost solution compared to traditional methods.

Keywords: Computer Vision; Logistics Operations; Convolutional Neural Networks; Object Detection; Industrial Security.

1 Introduction

With the advent of the popularization of the application of computational techniques involving artificial intelligence, there was the enhancement of applications in the most diverse areas of knowledge, being driven by the optimization of pre-existing techniques, reduction of process times and greater practicality and ease of implementation of advanced solutions.

The present study aims to reconcile these techniques that involve AI and heuristic concepts, pre-existing algorithms associated with image processing techniques, applying them in a customized way in logistics operations that require constant monitoring of points of interest, involving security aspects, minimizing the risks existing in operations where machines and pedestrians coexist in the same environment/scene.

The resources available on the market present points of vulnerability, where there is an opportunity to study specific techniques and integrations with the potential to positively contribute to the current state of the art, regarding actuation and active protection devices in logistics operations, involving mobile machines, which are interacting with pedestrians or other machines.

One of the main objectives of the study developed seeks to raise safety levels in risk areas in industrial environments and act as a tool that contributes to the reduction of accidents and consequently favoring the preservation of human life, acting indirectly in the preservation of assets and infrastructure existing in these areas since it aims to mitigate the risk of collisions.

2 Problem Characterization

The focus of the research differs from the solutions commonly applied for this purpose, which use invasive devices installed on the machines and equipment that are monitored, requiring the installation of multiple devices

(if there is more than one piece of equipment in the same area, each machine must be equipped with a set of several specific cameras and sensors).

In addition to requiring time-consuming installations, which require integration into the machines' native systems, with the risk of generating failures, loss of warranty and finally, embedded devices, such as cameras in mobile equipment, are more prone to failures due to excessive vibration, accumulation of dirt or even data resulting from misuse and changes in the positioning of accessories.

Computer vision-based systems are limited to replicating the image detection and identification model and acting in a similar way to human vision. In other words, for the object of interest to be identified, it needs to be within the viewing range of the camera or any other sensor.

Operations in industrial environments typically present layout and positioning conditions for machines and pedestrians that, under certain exposure conditions, can lead to risks of collision, physical and material damage, for example, in conditions where the machine and pedestrian are in opposite aisles and intersection points, as shown in the figure. 1, which illustrates the possible situation.

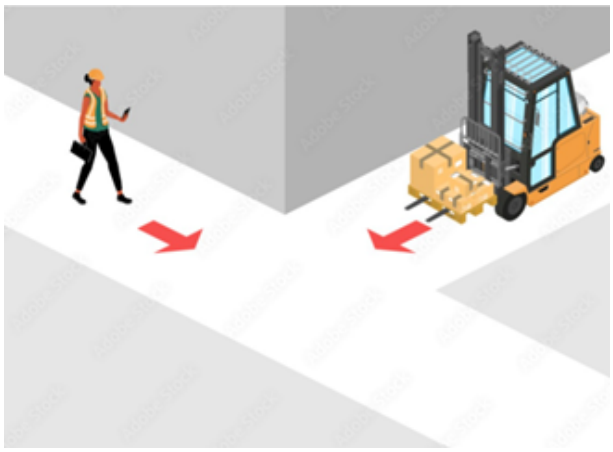


Figure 1. Example of pedestrian and machine interaction: Adobe Stock.

The scenario illustrated in the figure 1, demonstrates a common condition in industrial settings. In the example, we have a machine (forklift) and a pedestrian on a collision course where neither has a clear view of each other. Depending on the size of the mobile machine, its speed, and the general environmental conditions—be it lighting, noise, or any distractions—this could lead to a serious operational incident. Conventional systems based solely on vision are inefficient, as they must detect objects of interest, process, signal, and intervene between the agents in the scene. Depending on the proximity, this entire flow and the fractions of a second involved may not be sufficient for human intervention, and the risk of accidents is imminent.

2.1 Summary of the Literature Review

The type of technique that will be used for practical applications involving the concept of computer vision is directly related to the level of information that you want to extract from the scene, as well as the treatments that you want to carry out with the data after identification, making it possible to reach the level of real-time 3D detection of the position and gestures of the human body, according to a study developed by Shotton[1],study published by Microsoft that describes the conceptual bases applied in "Kinect", where images are captured in real time, decoded and directly linked to the application of interest, in this case electronic games.

Studies in the area of computer vision typically use a multidisciplinary approach, where techniques involving artificial intelligence and machine learning are combined, which are customized for the most diverse applications, be it in industry, the medical field, sports entertainment, autonomous vehicles, etc., according to a study developed by Karn [2].

Whatever the application to be developed, there are some common points: the need to capture the image, whether in real time or from a base of specific images, the training of the algorithm in light of the scenario of interest and the configuration of the "outputs", in light of the desired objective. Within this concept, the work developed by Kapoor can be seen. [3], where a methodology for identification through aerial image capture is presented, focusing on the identification of the human body within the scenario, even allowing the recognition of human actions (walking, strolling, sitting, etc.), to achieve the objective of the study, YOLO V4 was used for identification of people and convolutional networks (M-ResNeXt101), for identification and classification of human actions, with the base of images available on the web.

According to Barbado [4] satisfactory results were achieved with the use of video monitoring cameras installed at fixed points, for vehicle detection, in this work the algorithms were customized in order to obtain the following information: determination of the direction and flow of vehicle movement, the model presented an assertiveness rate in detection of 76 percent, to achieve the result presented the following libraries were used: OpenCV (Open Source Computer Vision Library), NumPy, MatPlotLib, CUDA Toolkit, and finally YOLO: Real-Time Object Detection (version 4), Using an NVidia GPU of the Tesla architecture, processing images above 60 Frames per Second (FPS).

Similar work was developed by Batista [5], with the objective of measuring the crossing time of pedestrians at road crossings, in order to allow the analysis of the delay time between the opening of the traffic light and crossing the road, with a focus on safety, with an impact on urban traffic management, for this purpose, yolo version 4 was used, from which adaptations were made in Python to collect the flows, delays and accepted gaps of pedestrians, the calculated metrics used a database of pre-recorded images as a reference.

In line with previous development, it is possible to identify works that follow a similar line of development, but using images taken in real time as an object of evaluation, as described by Cortez [6], where, through computer vision integrated with traffic light instruments, it was possible to increase traffic flow by 33 percent, using the vehicle flow count on the road as a reference, to act in real time during the traffic light opening and closing retention interval.

According to Jiaqi Geng [7] he states that the detection of points of interest in images, especially human silhouettes, is negatively influenced by the risk of image occlusion in the scene and possible lighting deficiencies. Radar and Lidar-based technologies, on the other hand, require specialized hardware, are expensive, and consume more energy. To address this problem, Geng proposed the development of a methodology that uses a machine learning model, typically used in computer vision, but uses input signals from Wi-Fi networks as a reference, mapping the phase and amplitude of these signals to obtain the position coordinates of the detected human forms, achieving results comparable to purely image-based approaches.

In the work developed by Moitinho [8],the application of machine learning is verified, associated with computer vision with the objective of identifying and tracking the position of people moving in an indoor environment, this article differs from others researched so far since it uses an ESP32-CAM board to capture images, connected to a camera, and the images are sent to an MQTT (Message Queue Telemetry Transport) broker, a microservices platform that connects to the components of the model developed for the application, in this case the solution is aimed at performing facial recognition, but provides bases for improving the model proposed in this article.

Considering the criticality of the topic, in addition to the aspects related to the visibility assistance algorithms that will be applied, the construction aspects of the mobile equipment that circulates through industrial areas must be taken into account, considering the standardized methods for evaluating these vehicles in relation to safety standards and existing regulations ISO 13564-1:2012 (Powered industrial forklifts: Test methods for verifying visibility) and ANSI/ITSDF B56.1-2020, which address these aspects as described by Bostelman and Teizer. [9].

2.2 Yolo V8

YOLO, or "You Only Look Once," is a popular framework for object detection in images and videos. It is widely used in computer vision applications such as video surveillance, autonomous vehicles, and object recognition. It was developed by Ultralytics. Version 8 was released in February 2023 and offers extensive support material available online for reference and development.

The developed model used YoloV8, a pre-trained computer vision model. Depending on the application and characteristics of the solution being developed, the model has five versions: Yolov8n (nano), Yolov8s (small), Yolov8m (medium), Yolov8l (large), and Yolov8x (extra large). The versions differ in size (requiring less computational effort to run) and speed of detecting the desired item. The Nano version (Yolo v8n) is the smallest and fastest, while the Extra Large version (Yolov8x) is more robust to run and therefore slower to detect.

As described in the figure 2, The decision to use the Yolo v8n model was based on the following points: The YOLOv8n model offers a combination of performance, consolidated robustness, and faster image processing combined with low response latency, essential requirements for industrial scenarios where real-time decision-making is critical.

Although newer models, such as YOLO11, are already available and bring improvements in accuracy and overall efficiency, independent tests and benchmarks point to YOLOv8n as the model capable of achieving superior inference speeds in real-world applications, especially in the "nano" (v8n) segment. According to detailed measurements, YOLOv8n infers an image in approximately 3.3 milliseconds, while its more recent counterpart,

Model	size (pixels)	mAP^{val} 50-95	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)	FLOPs (B)
YOLOv8n	640	37.3	80.4	0.99	3.2	8.7
YOLOv8s	640	44.9	128.4	1.20	11.2	28.6
YOLOv8m	640	50.2	234.7	1.83	25.9	78.9
YOLOv8l	640	52.9	375.2	2.39	43.7	165.2
YOLOv8x	640	53.9	479.1	3.53	68.2	257.8

Figure 2. Accuracy Yolo V8: Ultralytics [10]

YOLO11n, operates in around 4.8 milliseconds, as described in Sapkota and M. Karkee. [11]. This difference, although it may seem small, directly impacts applications where accumulated delays impair automatic operation, for example, when activating security devices in real time and remotely.

In addition to its speed, YOLOv8n boasts proven robustness, being widely documented in recent commercial and industrial applications, providing greater predictability and stability to the production cycle. YOLOv8n's low computational consumption also allows its deployment on affordable hardware, reducing costs and facilitating future maintenance. In latency-sensitive applications, such as this project, choosing YOLOv8n ensures that each detection made by the model is transmitted and processed almost instantaneously, contributing to the safety, efficiency, and reliability of the system as a whole.

2.3 Color Segmentation

The proposed model uses color segmentation in HSV space, a technique widely used in image processing due to its ability to separate color information (hue and saturation) from intensity (value), approaching human visual perception.

When converting images from RGB space to HSV, it is possible to segment objects more efficiently, adjusting thresholds in the H, S and V channels to isolate regions of interest, such as foreground objects or specific areas in images, this technique is used to limit regions in the monitored scene to which specific detection alerts will be triggered, if the areas with segmented colors are transported, Hema and Kannan. [12], present the development of a tool in Python, where color segmentation is used to extract fragments, aiming to improve the performance of detection algorithms.

HSV-based methods demonstrate better performance compared to other color spaces, facilitating object identification even under varying lighting conditions and allowing results closer to human perception, according to Zhang and Fan[13], propose a method that enhances low-light images with sharp details, improving contrast and reducing color distortion using semantic segmentation and HSV color space, outperforming traditional methods.

Computer vision algorithms are typically influenced by the level of efficiency depending on the lighting conditions in the scene where the object of interest is located, in the work developed by Wu, Wang and Huang,[14] propose an unsupervised image enhancement method based on the transformation of the HSV color space to improve the visual quality of low-light images in coal mines, where favorable results were observed, superior to algorithms based purely on computer vision. The studies presented corroborate the decision of the present project to use neural network detection, associated with color segmentation of the areas of interest.

2.4 Hardware Used

In line with the application's concept of efficiency and simplicity, the proposed model uses Wi-Fi communication, combining embedded systems with computer vision. For this purpose, an ESP32-WROOM 32 development board was used. The respective hardware features Wi-Fi communication over HTTP protocol, enabling device activation via IP, via remote web access. Thus, once the detection conditions are met, the computer vision algorithm sends the internal signaling signal to the vehicle being monitored by external cameras that view the environment. For functional testing, a 2K IP camera with a 3.6 mm lens, model A8 IC See, was used.

2.5 Methodology

The operating topology of the entire system consists of the following steps: reading data from the environment through an IP camera, accessing images via RTSP (Real Time Streaming Protocol) which allows control of real-time video transmission over the internet, processing images through a script developed in Python, which accesses the Yolov8n model that is installed on a Notebook connected to the Wifi network, after detection a command is sent to an ESP32 board, signaling that the object of interest has accessed near the area segmented by color.

The Yolov8n model was used, with a dataset of 1,500 images of people and vehicles, obtained from the Kaggle database, as identified Kshitij [15] e Sanagapati[16], The dataset was pre-processed in Roboflow using data augmentation (crop, zoom and duplication),resulting in a dataset of images for use in training with 3612 images, and a total of 2568 instances(1,352 annotated as people and 1,216 annotated as vehicles).

Image annotations for classification were performed manually using transfer learning, which was essential for the efficient training of the YOLOv8n model. Using Roboflow enabled the creation of a robust and standardized dataset, facilitating not only the labeling of regions of interest but also preprocessing and the division of the data into training, validation, and test sets, optimizing the workflow and the quality of the material used.

Using this dataset, the YOLO model, previously trained on large datasets, was fine-tuned to the specific context of the project. This approach reduced training time, required less labeled data, and provided greater accuracy, making the model more efficient and reliable for detecting critical situations in real time.

The model described was trained for 100 epochs. In parallel, the following developments were carried out: the code was implemented via Arduino IDE for the ESP32 development board, which receives HTTP feedback from the developed detection script. Images are captured by IP cameras via RTSP. Color segmentation is applied to delimit the monitored objects. Whenever detection by the Yolov8n computer vision algorithm occurs near these segmented areas, a specific alert is issued.

The entire system operates in a closed circuit that combines remote image processing and Wi-Fi transmission of information from detections made.

2.6 Results

Analyzing the results presented by the graphs generated by the Yolov8n model, we can see good identification and categorization of people and vehicles. The model's accuracy curve indicates consistent performance, with an overall accuracy level of 0.98. Evaluating the performance metrics, which take into account the progress of the indicators throughout the training process (100 epochs), we can observe a steady increase in the precision, recall, and MAP metrics. Specifically, the mAP50 indicator reaches levels between approximately 0.50 and 0.75, As shown in the figure 3, indicating that the model strikes a good balance between sensitivity and specificity, consistent with the work developed by Wang and Huawei [17]. The results presented enable the application of the model in real operational environments, performing adequately in low-light conditions and variable positioning conditions.

The satisfactory results presented were favored by the appropriate annotation of the image dataset, as well as the application of data augmentation techniques (crop, zoom, and duplication). Similar results were observed in the work developed by Panja, Hendry, and Dewl.l[18].

The proposed model presents results compatible with similar studies, used as a basis for the development of this article, presenting as differences in relation to previous studies, the proposed integrations including segmentation by color, which allows the sectorization of areas to be mapped, in addition to the architecture of remote detection and decentralized processing, with local report as described in figure 4.

The Model simplifies implementation and scalability as a viable solution for practical applications.In the figure 4, we have the illustration of the implemented operating topology, thus identified, in one (1), we visualize – the IP camera that captures the images, in two (2) we have the machine visualized by the camera, which can be a forklift or an area that we want to monitor, in three (3) we have the Notebook connected to the wifi network, where the Yolov8n model is processed, in four (4), we have a second machine (forklift), which receives the signaling, if the risk and proximity conditions described in the computer vision model are met.

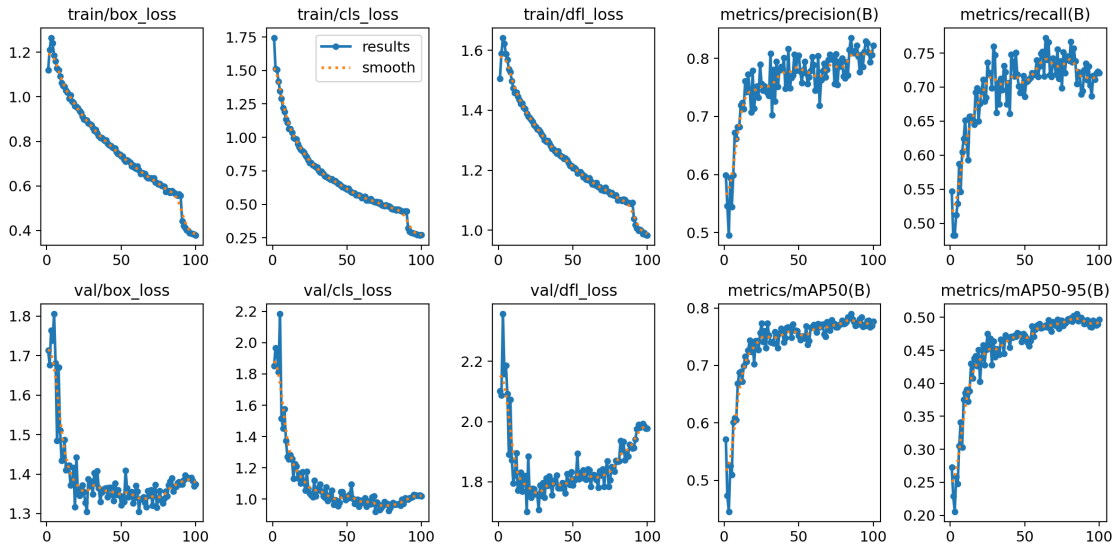


Figure 3. Results: Yolov8n

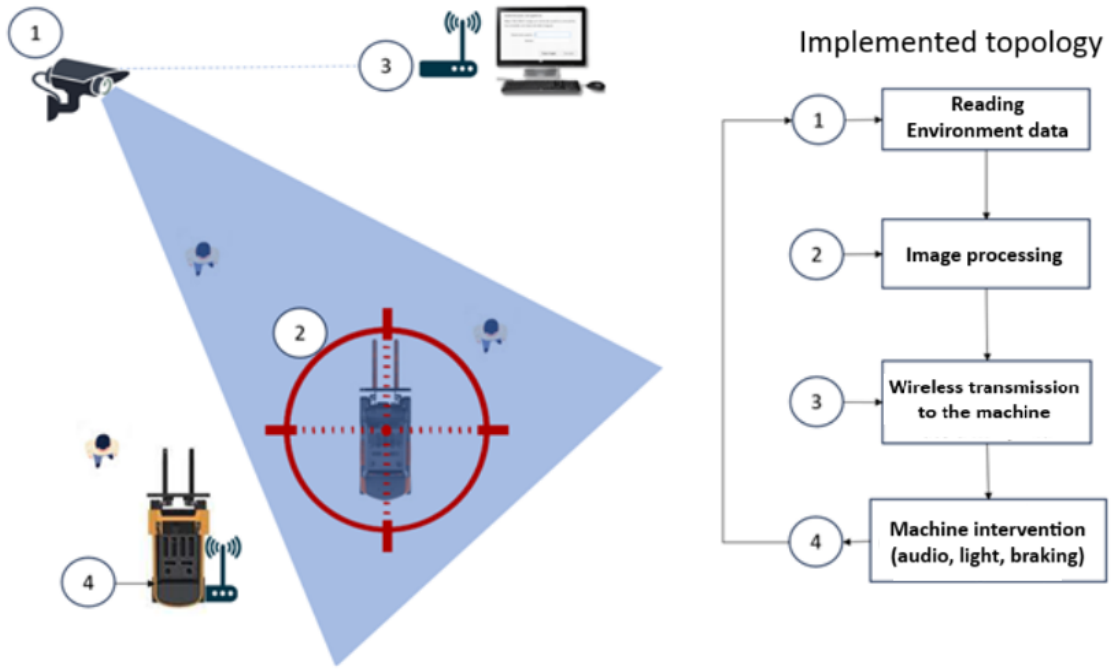


Figure 4. Illustration of the Operating Topology: Autor

3 Conclusion

The developed model fulfills its intended purpose, demonstrating the feasibility of applying the operational risk detection model to logistics environments using a decentralized development topology (remote processing), eliminating the need for embedded image processing hardware, cables, cameras, and complex installations that are more prone to failure. The combined use of color segmentation techniques and local alerts transmitted locally to objects of interest in the scene opens up a range of application possibilities that distinguish it from similar systems currently available on the market.

References

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pp. 1297–1304, 2011.
[2] A. Karn. Artificial intelligence in computer vision. *International Journal of Engineering Applied Sciences and Technology*, vol. 6, pp. 249–254, 2021.
[3] S. Kapoor, A. Sharma, A. Verma, and S. Singh. Aeriform in-action: A novel dataset for human action recognition in aerial videos. *Pattern Recognition*, vol. 140, pp. 109505, 2023.
[4] L. Barbado, L. Santos, M. Matrakas, and J. Moreira. Application of the yolo convolutional network for vehicle flow analysis. In *Proceedings of the XIX Latin American Congress on Free Software and Open Technologies*, pp. 43–49, Porto Alegre, RS, Brasil. SBC, 2022.
[5] Batista de F. A. Castro Junior, M. M. d. Castro Neto, and F. J. Craveiro Cunto. Analysis of pedestrian delay and accepted gap at signalized crossings: a study in the city of fortaleza using computer vision techniques based on deep learning. *TRANSPORTES*, vol. 31, n. 2, pp. e2845, 2023.
[6] D. CORTEZ. Development of an intelligent traffic control system based on computer vision. Master's thesis, FEDERAL UNIVERSITY OF RIO GRANDE DO NORTE, Natal, RN, 2022.
[7] J. Geng. Dense human pose estimation from wifi. Master's thesis, Carnegie Mellon University, Pittsburgh, PA, 2022.
[8] L. C. Moitinho and A. Benicasa. Machine learning to help locate people in indoor environments monitored by cameras. In *Extended Proceedings of the XIX Brazilian Symposium on Information Systems*, pp. 71–80, Porto Alegre, RS, Brasil. SBC, 2023.
[9] R. Bostelman, J. Teizer, S. J. Ray, M. Agronin, and D. Albanese. Methods for improving visibility measurement standards of powered industrial vehicles. *Safety Science*, vol. 62, pp. 257–270, 2014.
[10] G. Jocher, A. Chaurasia, and J. Qiu. *Ultralytics YOLOv8 Documentation*. Ultralytics. Accessed: 2025-10-11, 2023.
[11] R. Sapkota and M. Karkee. Comparing yolo11 and yolov8 for instance segmentation of occluded and non-occluded immature green fruits in complex orchard environment. *arXiv preprint*, vol. arXiv:2410.19869, 2024.
[12] D. Hema and D. S. Kannan. Interactive color image segmentation using hsv color space. *Science Technology Journal*, 2019.
[13] J. Y. e. C. F. Hang Zhang. Low-light image enhancement using semantic segmentation and hsv color space. *SSRN Electronic Journal*, 2022.
[14] Mine image enhancement based on hsv color space. *IEEE Access*, vol. 12, pp. 72170–72186, 2024.
[15] Kshitij. Cars image dataset. `https://www.kaggle.com/datasets/kshitij192/cars-image-dataset`. Acessado em: 13 set. 2025, 2025.
[16] P. Sanagapati. Images dataset. `https://www.kaggle.com/datasets/pavansanagapati/images-dataset`. Acessado em: 13 set. 2025, 2025.
[17] B. Wang, Y.-Y. Li, W. Xu, H. Wang, and L. Hu. Vehicle–pedestrian detection method based on improved yolov8. *Electronics*, 2024.
[18] E. Panja, H. Hendry, and C. Dewi. Yolov8 analysis for vehicle classification under various image conditions. *Scientific Journal of Informatics*, 2024.