1. Introduction

The statistics of motorized vehicles in Indonesia according to the Indonesian National Police Traffic Corps (*Korlantas Polri*) show that the motorized vehicles reached a staggering 160,652,675 in February of 2024, which has significantly increased from the previously recorded 148,261,817 in 2022. This represents a significant increase over the past two years. However, alongside this increase, the infrastructure necessary for the vehicles hasn't improved and the roads have not expanded. Due to this imbalance, the traffic conditions became unbearable which in turn resulted in higher traffic violations as well.

Among the traffic violations, zebra-cross violations are particularly concerning due to their direct impact on pedestrian safety[1]. Pedestrian crossings are often disregarded by drivers, leading to accidents and increasing the risk to vulnerable road users. For instance, according to data from *Korlantas Polri*, 8.274 traffic accidents occurred in 2023 involving pedestrians crossing the road. While the government has issued traffic regulations as regulated in Article 287 paragraph 1 of Law Number 22 of 2009 concerning Traffic and Road Transportation (LLAJ Law) concerning the obligation of drivers to obey command or prohibition signs and road markings, enforcement still remains a challenge, emphasizing the importance of technological solutions to detect and mitigate violations.

Advancements in computer vision and machine learning have enabled the development of automated methods for traffic monitoring and rule enforcement. The YOLO (You Only Look Once) algorithm has proven effective in detecting objects quickly and efficiently [2]. While previous studies have explored the use of YOLO for zebra-cross violation detection, these efforts were limited by dataset scope, real-world applicability, and the ability to handle complex scenarios, such as multiple-object detection and varying camera angles.

This research applies YOLO (You Only Look Once), a real-time object detection method that efficiently detects objects in a single pass, streamlining the computational process (Wang et al., 2021). YOLO interprets image data as a regression problem, using deep learning to generate bounding boxes, labels, and confidence scores for detected (Reis et al., 2023; Wang et al., 2024). It is widely used in applications such as traffic signal detection, pedestrian monitoring, and parking identification due to its speed and accuracy [6], [7], [8].

YOLOv9 offers a superior balance of speed and accuracy compared to other object detection algorithms, making it ideal for real-time applications like zebra-cross violation detection. Unlike Faster R-CNN, which uses a slower two-stage approach , YOLOv9's single-stage design predicts bounding boxes and classes simultaneously, enabling faster inference [9]. While SSD provides real-time performance, it struggles with small object detection, an area where YOLOv9 excels due to its advanced GELAN backbone [10], [11], [12]. Compared to RetinaNet, YOLOv9 maintains similar accuracy but achieves higher frame rates, making it more efficient for real-world scenarios. Additionally, YOLOv9's flexibility with model variants (e.g., YOLOv9-n, YOLOv9-s) allows users to optimize performance based on hardware capabilities, further enhancing its usability.

In this study, we focus on developing a zebra-cross violation detection method that can handle real-world scenarios more dynamically using YOLOv9. The YOLOv9's improved architecture offers significant improvements over earlier versions[13], including enhanced accuracy, faster detection speeds, and better handling of complex scenarios [14]. The YOLO algorithm is suitable for this research because YOLO performs object detection and classification in a single end-to-end drilled network, which allows for more efficient learning compared to other models that require multiple training stages (e.g., models like Mask R-CNN that require a region proposal stage [15]. YOLO is also known for its real-time prediction speed [16], [17]. The model processes the entire image in a single convolutional step, making it very fast compared to other models that use region proposal-based approaches or pixel-wise segmentation[18].