Telematika – Vol. 19, No. 2, Augut (2026) pp. 183-197 e-ISSN 2442-4528 | p-ISSN 1979-925X

Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemdiktisaintek, No.10/C/C3/DT.05.00/2025

# Robustness Evaluation of YOLOv8, YOLOv11, and YOLOv12 for Personal Protective Equipment Detection under Photometric Saturation Variations

**Zaid Romegar Mair[1,*], Rudi Heriansyah[2], Septa Cahyani[3], M. Ravensky Taro Danayaksa[4]**

[1,3,4] Informatics Engineering, Faculty of Computer and Natural Science, Indo Global Mandiri University, Palembang, Indonesia

[2] Master of Computer Science, Faculty of Computer and Natural Science, Indo Global Mandiri University, Palembang, Indonesia

**ARTICLE INFO**

***History of the article:***
Received June 2, 2026
Revised July 26, 2026
Accepted August 1, 2026

**Keywords:**
Personal Protective Equipment
Photometric Saturation
YOLOv8
YOLOv11
YOLOv12

**Correspondence:**
zaidromegar@uigm.ac.id

**ABSTRACT**

Computer vision-based Personal Protective Equipment (PPE) detection has become increasingly important for improving Occupational Safety and Health (OSH) compliance through automated, real-time monitoring of workers' safety practices. Although recent studies have reported promising performance for YOLO-based object detection models, most evaluations have been conducted under standard imaging conditions, providing limited evidence of model robustness against photometric variations. To address this gap, this study proposes a systematic robustness evaluation framework based on controlled photometric saturation variations to compare the performance stability of three state-of-the-art object detection models: YOLOv8, YOLOv11, and YOLOv12, for multi-class PPE detection. The experimental dataset comprised 3,116 images annotated into eight PPE-related classes: Helmet On, No Helmet, Vest On, No Vest, Gloves On, No Gloves, Boots On, and No Boots. To ensure a fair comparison, all models were trained using identical experimental settings and evaluated using Precision, Recall, mAP@50, and mAP@50–95. Model robustness was assessed under three saturation conditions (−30%, 0%, and +30%), representing realistic color variations commonly encountered in construction-site surveillance. The experimental results revealed that photometric saturation variations produced only marginal changes in detection performance across all evaluated models. Among the three architectures, YOLOv8 achieved the highest overall performance, attaining an mAP@50 of 55.1%, compared with 52.9% for YOLOv11 and 48.6% for YOLOv12, while maintaining the most stable performance under varying saturation levels. Although YOLOv12 demonstrated relatively better capability for detecting several small-object classes, its overall detection performance remained inferior to that of YOLOv8. These findings indicate that YOLOv8 provides the best trade-off between detection accuracy and robustness under moderate photometric saturation variations. This study contributes a systematic robustness evaluation of recent YOLO architectures under controlled photometric conditions and offers practical insights for selecting reliable object detection models for real-world PPE monitoring systems.

## 1. INTRODUCTION

Occupational Safety and Health (OSH) has become a major concern in construction and industrial sectors due to the high incidence of workplace accidents. Proper use of Personal Protective Equipment

(PPE), including safety helmets, reflective vests, gloves, and safety boots, plays a crucial role in reducing occupational injuries and ensuring compliance with established safety regulations. Consequently, reliable monitoring of PPE usage has become an essential component of modern occupational safety management systems. (Laily et al., 2022; Santos et al., 2025).

Despite established safety regulations, PPE compliance remains inconsistent due to worker negligence, discomfort, and limited supervision, reducing the effectiveness of conventional manual monitoring systems (Mailoa & Santoso, 2022). Although recent studies have investigated robustness in object detection under illumination changes, motion blur, adverse weather, and image noise, comparatively fewer studies have examined photometric saturation variation, particularly for multi-class PPE detection using recent YOLO architectures. Since surveillance images acquired from CCTV cameras and edge devices frequently experience color saturation changes caused by illumination, camera calibration, and automatic white balance, evaluating model robustness under controlled saturation variation is essential for reliable real-world deployment (Ofori-Oduro & Amer, 2024; Gholinavaz et al., 2025).

Recent advancements in deep learning, particularly object detection frameworks such as YOLOv5 and YOLOv8, have shown significant potential in enhancing real-time surveillance for personal protective equipment (PPE) compliance in construction and industrial environments. For instance, (Nazli et al., 2024) developed a real-time PPE detection system using YOLOv5 that achieved an impressive mAP of 91.9% and 98.3% accuracy, effectively classifying workers' safety levels. Similarly, (Kim & Xiong, 2025) implemented an improved YOLOv8-based model on edge devices to monitor PPE usage, contributing to proactive accident prevention, especially falls from heights. Unlike conventional benchmarking studies that compare object detection models only based on accuracy under standard imaging conditions, this study emphasizes robustness evaluation by systematically measuring detection performance under controlled photometric saturation variations. Such evaluation provides additional evidence regarding model stability and reliability when visual characteristics change while object semantics remain unchanged. Therefore, this study contributes not only to comparative model evaluation but also to the understanding of robustness characteristics required for reliable deployment in real-world occupational safety monitoring systems (Gyawali et al., 2026; Gholinavaz et al., 2025).

Convolutional Neural Networks (CNNs) have become one of the most effective deep learning approaches for object detection owing to their capability to learn hierarchical visual representations from image data, especially in the form of the You Only Look Once (YOLO) framework. YOLO is known for its ability to detect various objects in a single inference process, making it very suitable for use in real-time monitoring systems in work environments (Ramos & Sappa, 2025; Cai et al., 2025). The application of YOLO in the context of OSH has been carried out in various studies, as shown in (Mailoa & Santoso, 2022), while (Wang et al., 2023) demonstrated that an improved YOLOX model combined with low-light image enhancement can improve PPE detection performance in poor illumination conditions. Despite these advances, most previous studies only used one version of YOLO and limited objects, thus not providing a comprehensive picture of the effectiveness of the latest versions in detecting various safety equipment. As newer versions of YOLO, such as YOLOv8, YOLOv11, and YOLOv12, continue to develop, research is needed to examine the performance of these three versions in the same context.

YOLOv8 comes with improved efficiency and accuracy through architecture optimization and augmentation pipeline (Irawan, 2024; Jocher et al., 2023), while YOLOv12 offers further enhancements in segmentation and small object recognition (Khanam & Hussain, 2025). Moreover, most studies have not tested the model's resilience to visual changes in images, such as changes in color saturation, even though this is an important factor that commonly occurs in the field due to lighting, camera quality, or environmental effects. Accordingly, this study further evaluates using controlled photometric saturation variations in three variants, namely low (-30%) and high (30%), to assess the robustness of the YOLOv8, YOLOv11, and YOLOv12 models in detecting PPE objects under varying and extreme lighting conditions (De & Pedersen, 2021). Photometric saturation variation represents one of the most common visual disturbances encountered in practical construction environments. In real-world surveillance systems, images acquired from different CCTV cameras, mobile devices, and edge-based monitoring systems are frequently affected by automatic white balance adjustment, color enhancement algorithms, sensor characteristics, illumination intensity, weather conditions, and camera calibration. These factors alter image saturation without changing the physical appearance of PPE objects, potentially influencing the feature representation learned by deep learning models. Consequently, evaluating object detection performance only under standard image conditions does not adequately reflect deployment scenarios encountered in industrial safety monitoring (Ofori-Oduro & Amer, 2024).

In this context, image enhancement techniques such as histogram equalization and multiscale fusion have been effectively used to improve image quality under challenging conditions, including underwater and low-light environment (Cahyani et al., 2024). Such preprocessing methods may offer complementary value to detection systems by optimizing visual clarity before inference.

Recent studies have emphasized the influence of both environmental conditions and computational configurations on the performance of object detection models for PPE compliance. For instance, (Ahmed et al., 2023) demonstrated that hardware constraints and lighting variability can substantially affect detection reliability, highlighting the importance of robust models. Comparative analyses by (Elesawy et al., 2024; Biswas et al., 2023) show that newer YOLO versions, such as YOLOv5 and YOLOv8, significantly outperform earlier versions in terms of mAP and speed. However, few studies have systematically evaluated these models under challenging, real-world construction conditions with diverse PPE classes. Table 1 presents a research gap analysis of previous studies related to PPE detection using YOLO-based architectures.

**Table 1.** Research Gap Analysis of Previous Studies

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | This Study |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PPE Detection | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ |
| Multi PPE Detection | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ |
| YOLOv5 | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ |
| YOLOv8 | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |
| YOLOv11 | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| YOLOv12 | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Comparative YOLO Evaluation | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ |
| Real Time Detection | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Robustness Analysis | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ◐ | ✓ | ✓ | ✓ | ✓ |
| Photometric Saturation Testing | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Evaluation under Visual Variations | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ |

Note : ✓ = Supported, ✗ = Not supported, ◐ = Partially supported

Various studies have proven that deep learning-based approaches, particularly the YOLO algorithm, have high potential in the application of automatic personal protective equipment (PPE) detection systems in construction work environments. This need is driven by the requirement for a real-time monitoring system capable of assessing workers' compliance with safety standards in workplace environments (Xiong & Tang, 2021).

Recent advances in YOLO-based PPE detection have focused on improving detection accuracy through architectural enhancements, including attention mechanisms, multiscale feature fusion, and improved localization loss functions (Kisaezehra et al., 2023; Li et al., 2022; Ruaz, 2023). Despite these improvements, previous studies have primarily emphasized detection accuracy rather than robustness under photometric variations. Table 2 summarizes representative studies and highlights the remaining research gaps addressed in this work.

**Table 2.** Benchmarking PPE Detection

| Study | Model | Dataset Size | PPE Classes | mAP (%) | Main Contribution | Limitation |
|---|---|---|---|---|---|---|
| Mailoa and Santoso (2022) | YOLO + ResNet50 | 2,738 images | 2 class Helmet, Vest | 96% Accuracy | Hybrid YOLO CNN classification | Only 2 classes of PPE. |
| Ruaz (2023) | Improved YOLOv3 | Not Reported,web crawler images | 2 Classes Helmet, No Helmet | 96.63 mAP50 | Improved YOLOv3 with multi-scale training for accurate and real-time helmet detection | did not investigate robustness under photometric variations or visual quality degradation. |
| Ahmed et al. (2023) | Fast R-CNN | 1,699 images | 8 classes | 96.0 mAP | Multi color helmet and PPE detection | Not testing visual variations. |
| Kisaezehra et al. (2023) | YOLOv5x | 7,063 images | 1 class Helmet | 95.8 mAP | High speed helmet detection | Helmet detection only. |
| Nazli et al. (2024) | YOLOv5 | 1,129 images | 7 classes | 91.5 mAP | Web based PPE compliance monitoring | Does not discuss visual degradation. |
| Elesawy et al. (2024) | YOLOv5-v8 | 1,330 images | Person, Vest, Helmet Colors | 92.3 mAP | Large scale YOLO comparison | No photometric robustness evaluation. |
| Kim and Xiong (2025) | Improved YOLOv8 Edge Device | 12,000 images | 3 class Helmet, Harness, Lanyard | 92.52 mAP50 | Coordinate Attention, Ghost Conv Transfer Learning, Merge-NMS on Jetson Xavier NX | Focus on fall-protection PPE, not photometric robustness. |
| Gholinavaz et al. (2025) | YOLOv5s, YOLOv8m, YOLOv10n, Faster R-CNN | 1,000 images (DAWN) | Traffic Objects | YOLOv8m best baseline | Comprehensive robustness evaluation under real weather and synthetic noise | Not designed for PPE detection and does not evaluate photometric saturation. |
| Gyawali et al. (2026) | YOLOv8n | 6,650 images | Helmet, Vest | 89.2 mAP50 | Physics-based weather augmentation for robust PPE detection under adverse environmental conditions | Does not evaluate photometric saturation variation and compares only one YOLO architecture. |
| Ofori-Oduro and Amer (2024) | YOLOv4 (also evaluated other detectors) | COCO & PASCAL VOC | General Objects | +9.19% robustness improvement | Robust object detection against multiple image distortions using KDE-based data augmentation | General object detection only; no PPE dataset and no photometric saturation evaluation. |
| Present Study | YOLOv8, YOLOv11, YOLOv12 | 3,116 images | 8 Classes | 55.1 / 52.9 / 48.6 mAP50 | Photometric saturation robustness evaluation (-30%, 0%, +30%) and comparison of YOLOv8-v11-v12 | Lower detection accuracy. |

To address these research gaps, the present study systematically evaluates the robustness of YOLOv8, YOLOv11, and YOLOv12 for multi-class PPE detection under controlled photometric saturation variations. The primary novelty of this work lies in the first systematic robustness evaluation of YOLOv8, YOLOv11, and YOLOv12 under controlled photometric saturation variations for multi-class PPE detection. Unlike previous studies that primarily focused on detection accuracy, this work investigates model stability under realistic visual disturbances while maintaining identical experimental settings across all evaluated architectures.

## 2. RESEARCH METHODS

This research adopts an experimental approach to compare the performance of YOLOv8, YOLOv11, and YOLOv12 under different saturation conditions for Personal Protective Equipment (PPE) detection. Similar comparative studies have demonstrated the effectiveness of evaluating multiple YOLO architectures for PPE compliance monitoring in construction environments (Sivanraj et al., 2025).

### 2.1. Research Flow

Figure 1 presents the overall workflow of the proposed study. The methodology comprises five sequential stages: (1) dataset collection and annotation, (2) image preprocessing and photometric saturation augmentation, (3) model training using YOLOv8, YOLOv11, and YOLOv12 under identical experimental settings, (4) performance evaluation using Precision, Recall, mAP@50, and mAP@50–95, and (5) comparative robustness analysis across three photometric saturation levels (−30%, 0%, and +30%). Each stage is described in the following subsections.

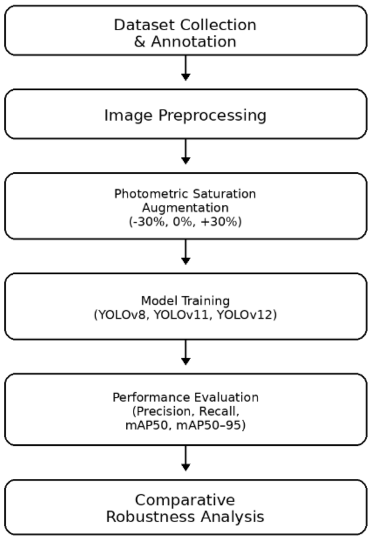



**Figure 1.** Research Flowchart

### 2.2. Data Collection and Annotation Data

The dataset used in this work consisted of 1,072 original images, which were collected from two sources. The training dataset was obtained from the public Roboflow repository, whereas the testing dataset comprised images directly documented from the construction work environment of the Omar Group. Before augmentation, 1,022 images were allocated for model training, while 50 images were reserved as an independent testing dataset. Data augmentation was applied exclusively to the training dataset, increasing the number of training images from 1,022 to 3,066. Consequently, the total dataset used in this study comprised 3,116 images, including the augmented training images and the original testing images. The augmentation strategy was employed to increase the diversity of training samples and improve model generalization, while the testing dataset remained unchanged to ensure an unbiased evaluation of model performance.

The dataset was annotated using a web-based annotation tool that generated labels in the YOLO format. A total of eight PPE-related classes were considered, namely Helmet On, Vest On, Gloves On, Boots On, No Helmet, No Vest, No Gloves, and No Boots. As shown in table 3, the object distribution among the eight classes was imbalanced. However, the same dataset partition, annotation files, and training configuration were consistently used for all three YOLO architectures (YOLOv8, YOLOv11, and YOLOv12) to ensure a fair comparative evaluation. No class weighting, oversampling, or resampling strategy was applied because the primary objective of the proposed evaluation was to evaluate the robustness of different YOLO architectures under identical experimental conditions rather than to maximize absolute detection accuracy.

### 2.3. Preprocessing and Augmentation

Data preprocessing and augmentation were performed automatically using the Roboflow cloud-based platform to increase the diversity of the training dataset without manually collecting additional images. All images first underwent preprocessing, including auto-orientation and resizing to 640 × 640 pixels, ensuring consistency with the input resolution required by the YOLO architectures. Subsequently, two categories of augmentation were applied. The first category was geometric augmentation, including rotation, flipping, cropping, and blurring, which simulated variations in camera viewpoint, object orientation, and image quality. The second category was photometric augmentation, including adjustments to hue, saturation, brightness, and exposure, to simulate variations in illumination and camera imaging conditions.

To evaluate the robustness of YOLOv8, YOLOv11, and YOLOv12 under photometric variations, this study focused on saturation adjustment using three levels: −30%, 0% (original image), and +30%. These values were selected because they represent moderate yet realistic color variations commonly encountered in CCTV surveillance systems, mobile cameras, automatic white balance correction, and illumination changes while preserving object visibility.

Conceptually, the saturation transformation was performed in the HSV (Hue–Saturation–Value) color space. Each RGB image was first converted into the HSV representation, where only the Saturation (S) channel was modified while the Hue (H) and Value (V) channels remained unchanged. After the saturation adjustment, the image was converted back to the RGB color space before being used for model training. The conceptual saturation adjustment is expressed as

$$S' = S \times (1+\alpha) \tag{1}$$

Where S is the original saturation value, $S'$ is the adjusted saturation value, and $\alpha$ represents the saturation factor with values of −0.3, 0, and +0.3. This equation is presented solely to illustrate the principle of saturation modification; the actual preprocessing and augmentation were performed automatically using the Roboflow augmentation pipeline.

All augmentation processes were applied before model training and only to the training dataset, whereas the testing dataset remained unchanged to ensure an unbiased evaluation of model robustness. The Roboflow platform automatically generated the corresponding annotation files by updating the bounding boxes for every augmented image (Roboflow, 2025).

### 2.4. Implementation of YOLOv8, YOLOv11, and YOLOv12

Data training was conducted using three lightweight YOLO architectures, namely YOLOv8n, YOLOv11n, and YOLOv12n, which represent the nano variants of their respective YOLO families. The nano variants were selected because they provide a favorable balance between computational efficiency and detection accuracy, making them suitable for real-time Personal Protective Equipment (PPE) monitoring while enabling a fair comparative evaluation under identical experimental conditions.

To ensure a fair and objective comparison, all models were trained using an identical experimental configuration. The standardized training parameters included 60 epochs, a batch size of 16, an input image resolution of 640 × 640 pixels, the Ultralytics automatic optimizer (optimizer = auto), an initial learning rate of 0.01, a final learning rate factor of 0.01, momentum of 0.937, weight decay of 0.0005, a patience value of 100 epochs, and a deterministic training setting with a fixed random seed of 0. These parameters were consistently applied across all experiments to eliminate the influence of hyperparameter variation and ensure that the observed performance differences were attributable primarily to the model architectures rather than differences in training configuration.

A training duration of 60 epochs was selected to provide sufficient optimization while maintaining computational efficiency, whereas a batch size of 16 was adopted as an appropriate compromise between GPU memory limitations and training stability. All experiments were conducted using the Google Colab cloud computing environment. The computational platform consisted of Python 3.12.13, PyTorch 2.11.0+cu128, and Ultralytics YOLO version 8.4.107, accelerated by an NVIDIA Tesla T4 GPU (14,913 MiB) with CUDA 12.8. The runtime environment also provided 2 CPU cores, 12.7 GB RAM, and 112.6 GB of available disk storage. The same dataset partition and identical training configuration were consistently applied to all three YOLO architectures throughout the experiments to ensure an objective and reproducible comparative evaluation.

## 3. RESULTS AND DISCUSSION

### 3.1. Model Testing Configuration

The training process for all YOLO models was carried out using uniform parameter settings, as presented in Table 2, which includes 60 training epochs, a batch size of 16, a default learning rate, and an input image resolution of 640×640 pixels. These configurations were applied consistently across all tested models to ensure fair performance comparison. The results presented in Table 3 correspond to the detection performance of each object class before the saturation enhancement was applied. The table to indicate the best values, intermediate values, and the lowest values across the YOLO testing metrics.

**Table 3.** Testing Configuration Of YOLOv8, YOLOv11, And YOLOv12 Models.

| Object Class | Total | YOLOv8 | | | | YOLOv11 | | | | YOLOv12 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Pr | Re | m50 | m50-95 | Pr | Re | m50 | m50-95 | Pr | Re | m50 | m50-95 |
| Helmet On | 64 | 0,735 | 0,797 | 0,873 | 0,554 | 0,876 | 0,859 | 0,895 | 0,543 | 0,482 | 0,844 | 0,828 | 0,539 |
| NoHelmet | 19 | 0,599 | 0,368 | 0,413 | 0,266 | 0,838 | 0,547 | 0,640 | 0,400 | 0,451 | 0,368 | 0,477 | 0,306 |
| Vest On | 78 | 0,808 | 0,872 | 0,902 | 0,573 | 0,782 | 0,783 | 0,839 | 0,525 | 0,595 | 0,846 | 0,764 | 0,469 |
| No Vest | 5 | 1,000 | 0,749 | 0,962 | 0,634 | 0,830 | 0,400 | 0,579 | 0,438 | 0,864 | 0,400 | 0,570 | 0,350 |
| Gloves On | 9 | 0,100 | 0,222 | 0,107 | 0,075 | 0,277 | 0,444 | 0,253 | 0,165 | 0,064 | 0,333 | 0,147 | 0,108 |
| No Gloves | 76 | 0,312 | 0,474 | 0,353 | 0,130 | 0,492 | 0,368 | 0,351 | 0,132 | 0,260 | 0,579 | 0,398 | 0,155 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Boots On | 86 | 0,392 | 0,244 | 0,247 | 0,083 | 0,257 | 0,140 | 0,143 | 0,055 | 0,320 | 0,302 | 0,218 | 0,074 |
| Sum/Avg | 337 | 0,564 | 0,532 | 0,551 | 0,331 | 0,622 | 0,506 | 0,529 | 0,322 | 0,434 | 0,525 | 0,486 | 0,286 |

High saturation level settings with a maximum percentage of 30% are shown in the table below. Table 4 presents a detailed comparison of detection performance across different object classes under +30% saturation, evaluated using Precision, Recall, mAP@50, and mAP@50–95.

**Table 4.** Detection Performance Comparison of YOLO Models

| Model | YOLOv8 | YOLOv11 | YOLOv12 |
|---|---|---|---|
| Best of all 3 | 18 | 13 | 5 |
| Middle of all 3 | 7 | 13 | 9 |
| Worst of all 3 | 7 | 6 | 18 |

There are 13, 15, and 4 object classes with the best metric values for YOLOv8, YOLOv11, and YOLOv12, respectively. After applying saturation enhancement, the number of object classes with optimal detection performance increases, particularly in YOLOv11. This indicates that saturation plays a significant role in varying lighting and saturation conditions are show in Table 5.

**Table 5.** Object Detection Results With 30% Increased Saturation

| Object Class | Total | YOLOv8 | | | | YOLOv11 | | | | YOLOv12 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Pr | Re | m50 | m50-95 | Pr | Re | m50 | m50-95 | Pr | Re | m50 | m50-95 |
| Helmet On | 64 | 0,735 | 0,844 | 0,882 | 0,556 | 0,859 | 0,875 | 0,900 | 0,536 | 0,466 | 0,844 | 0,829 | 0,541 |
| NoHelmet | 19 | 0,591 | 0,368 | 0,398 | 0,266 | 0,768 | 0,579 | 0,594 | 0,393 | 0,411 | 0,316 | 0,473 | 0,306 |
| Vest On | 78 | 0,796 | 0,885 | 0,912 | 0,589 | 0,847 | 0,400 | 0,577 | 0,436 | 0,578 | 0,859 | 0,766 | 0,486 |
| No Vest | 5 | 1,000 | 0,896 | 0,995 | 0,647 | 0,746 | 0,795 | 0,838 | 0,532 | 0,676 | 0,422 | 0,576 | 0,329 |
| Gloves On | 9 | 0,109 | 0,333 | 0,176 | 0,125 | 0,254 | 0,444 | 0,231 | 0,150 | 0,023 | 0,111 | 0,139 | 0,091 |
| No Gloves | 76 | 0,303 | 0,461 | 0,295 | 0,112 | 0,471 | 0,329 | 0,357 | 0,129 | 0,244 | 0,509 | 0,346 | 0,142 |
| Boots On | 86 | 0,339 | 0,215 | 0,240 | 0,084 | 0,238 | 0,140 | 0,142 | 0,057 | 0,360 | 0,326 | 0,231 | 0,078 |
| Sum/Avg | 337 | 0,553 | 0,572 | 0,557 | 0,340 | 0,597 | 0,509 | 0,520 | 0,319 | 0,394 | 0,484 | 0,480 | 0,282 |

Meanwhile, the low saturation level settings with a maximum percentage of -30% are shown in Table 6.

**Table 6.** Detection Performance Comparison of YOLO Models After Saturation Enhancement (+30%).

| Model | YOLOv8 | YOLOv11 | YOLOv12 |
|---|---|---|---|
| Best of all 3 | 13 | 15 | 4 |
| Middle of all 3 | 16 | 9 | 7 |
| Worst of all 3 | 3 | 8 | 21 |

From the -30% saturation table, it can be seen that the YOLOv8 testing metric values still dominate are show in Table 7.

**Table 7.** Object Detection Results With 30% Decreased Saturation.

| Object Class | Total | YOLOv8 | | | | YOLOv11 | | | | YOLOv12 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Pr | Re | m50 | m50-95 | Pr | Re | m50 | m50-95 | Pr | Re | m50 | m50-95 |
| Helmet On | 64 | 0,736 | 0,781 | 0,853 | 0,537 | 0,758 | 0,844 | 0,847 | 0,494 | 0,498 | 0,812 | 0,814 | 0,518 |
| NoHelmet | 19 | 0,582 | 0,368 | 0,440 | 0,285 | 0,704 | 0,579 | 0,617 | 0,382 | 0,447 | 0,368 | 0,487 | 0,316 |
| Vest On | 78 | 0,792 | 0,821 | 0,867 | 0,568 | 0,721 | 0,846 | 0,826 | 0,530 | 0,611 | 0,821 | 0,761 | 0,463 |
| No Vest | 5 | 1,000 | 0,650 | 0,920 | 0,557 | 0,725 | 0,540 | 0,585 | 0,443 | 0,600 | 0,320 | 0,518 | 0,333 |
| Gloves On | 9 | 0,128 | 0,333 | 0,087 | 0,061 | 0,148 | 0,444 | 0,221 | 0,133 | 0,064 | 0,333 | 0,151 | 0,112 |
| No Gloves | 76 | 0,299 | 0,500 | 0,369 | 0,139 | 0,346 | 0,461 | 0,345 | 0,132 | 0,291 | 0,592 | 0,418 | 0,153 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Boots On | 86 | 0,384 | 0,256 | 0,252 | 0,083 | 0,227 | 0,209 | 0,141 | 0,054 | 0,372 | 0,326 | 0,229 | 0,079 |
| Sum/Avg | 337 | 0,560 | 0,530 | 0,541 | 0,319 | 0,519 | 0,561 | 0,512 | 0,310 | 0,412 | 0,510 | 0,483 | 0,282 |

From the presented tabular testing data, it can also be transformed into a graphical comparison. The YOLOv8 model consistently shows the best performance across all metrics (Precision, Recall, mAP50, and mAP50-95) compared to YOLOv11 and YOLOv12. Increasing saturation by up to (+30%) improves the model's performance across all metrics while decreasing saturation by (-30%).

**Table 8.** Detection Performance Comparison of YOLO Models After Saturation Enhancement (-30%).

| Model | YOLOv8 | YOLOv11 | YOLOv12 |
|---|---|---|---|
| Best of all 3 | 15 | 11 | 4 |
| Middle of all 3 | 11 | 13 | 12 |
| Worst of all 3 | 6 | 8 | 16 |

However, the condition where the image is not subjected to saturation changes (Original) yields stable results. Saturation is an important factor that affects the performance of the object detection model. Increasing saturation can enhance the model's ability to detect objects more effectively are shown in Figure. 3.

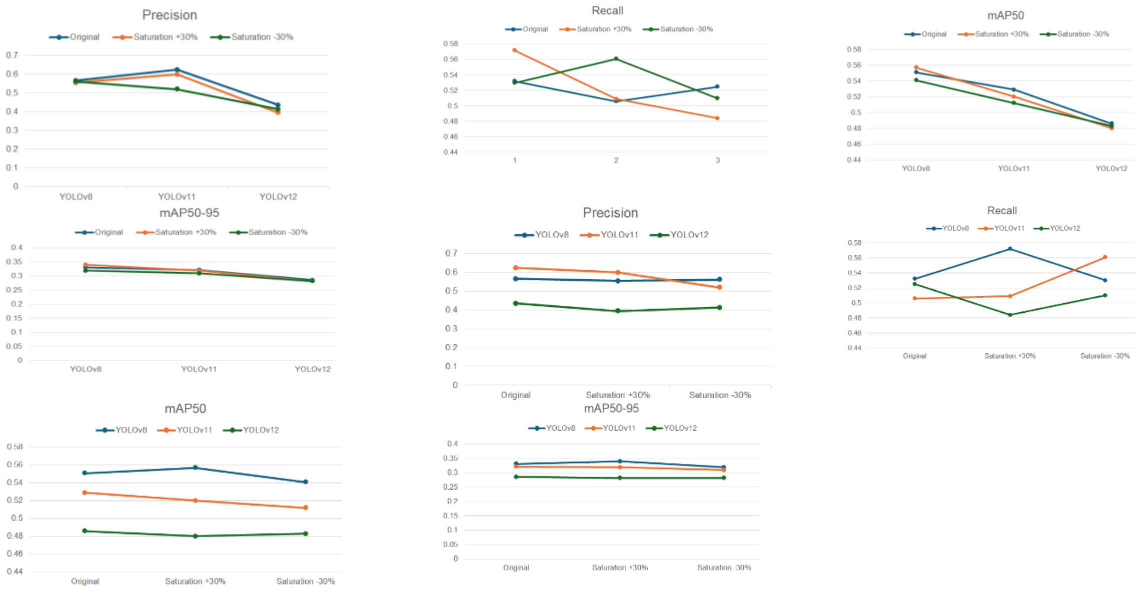



**Figure 3.** Saturation Comparison Diagram with YOLOv8, YOLOv11, and YOLOv12 Architectures

As shown in Figure 3, YOLOv8 consistently achieved the highest overall performance across all evaluated saturation levels, followed by YOLOv11 and YOLOv12. The superior performance of YOLOv8 can be attributed to its balanced detection architecture and more stable feature representation, which enabled consistent object detection despite moderate photometric saturation variations. In contrast, although YOLOv11 and YOLOv12 incorporate more recent architectural refinements to enhance feature learning and computational efficiency, these improvements did not result in superior overall performance under the identical dataset and training configuration adopted in this study. Because all models were trained and evaluated using the same experimental settings, the observed performance differences primarily reflect the interaction between model architecture and dataset characteristics rather than differences in the training strategy. These findings indicate that architectural complexity alone does not necessarily lead to improved robustness under controlled photometric saturation variations.

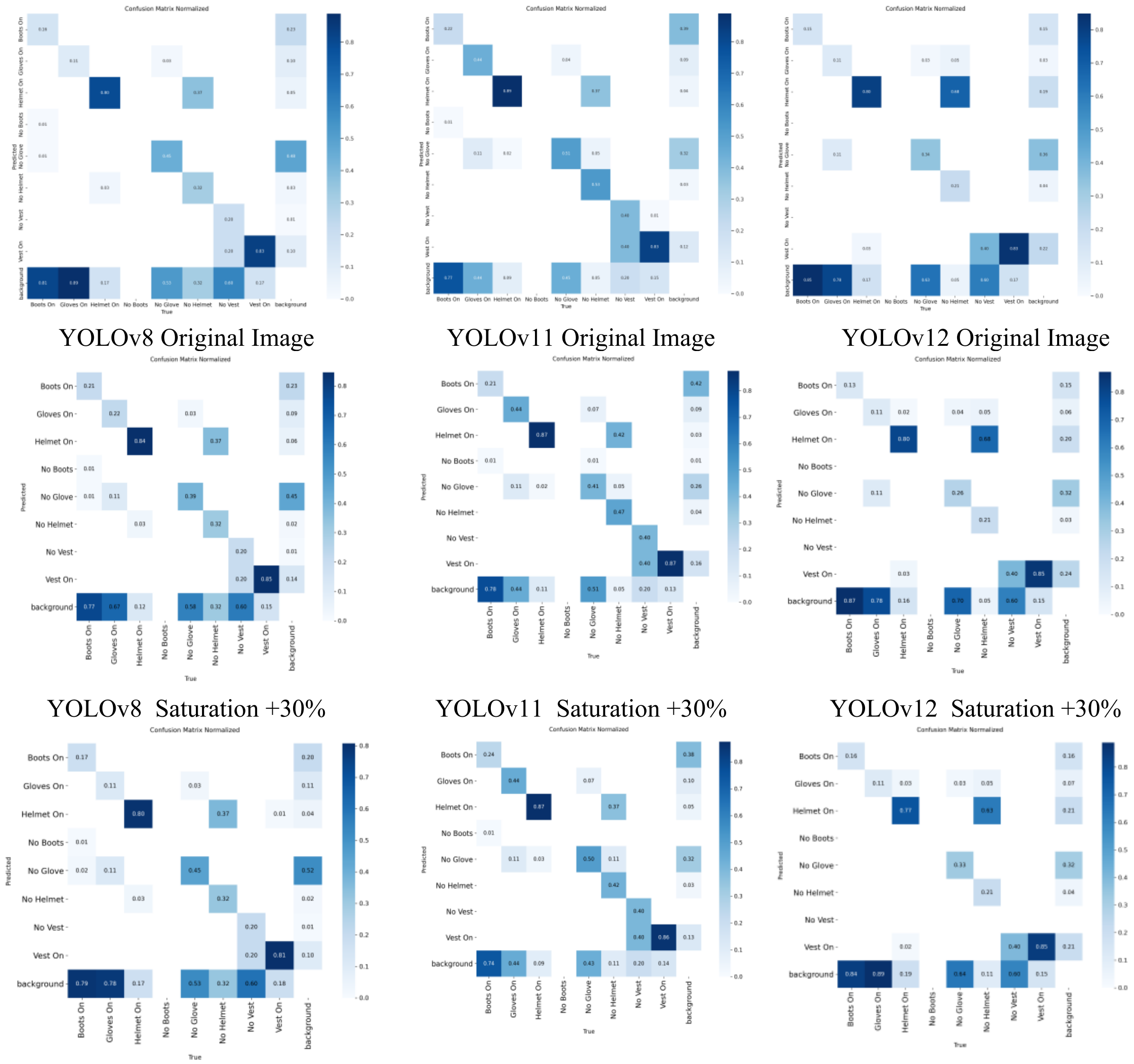
The experimental results are consistent with previous studies reporting that object detection performance is influenced not only by architectural design but also by dataset characteristics and application-specific conditions. More importantly, this study extends previous PPE detection research by demonstrating that robustness evaluation under controlled photometric saturation variations provides additional insights beyond conventional accuracy-based benchmarking. Consequently, comparative robustness assessment should be considered an essential component when selecting object detection models for real-world occupational safety monitoring systems, where variations in illumination, camera characteristics, and image quality are unavoidable.

### 3.2. Evaluation Model With A Confusion Matrix

The trained models are evaluated using quantitative metrics that include Precision, Recall, mAP@50, and mAP@50–95. These metrics are used to measure the precision and coverage of object detection by each version of YOLO. Additionally, the prediction results are also visualized in the form of a confusion matrix to assess the classification accuracy between classes. In the image matrix Figure 4., there are calculation results with the confusion matrix, and saturation changes of +30% and -30% for each YOLO version.

Further analysis of the confusion matrix reveals that most detection errors originated from false negatives, where PPE objects were classified as the background class rather than being confused with other PPE categories. This behavior was particularly evident for Boots On, with 70 false negatives, followed by No Glove (40), Helmet On (11), and Gloves On (8), indicating that the detector was more likely to miss PPE objects than incorrectly classify them into another PPE category. False positives were comparatively less frequent and primarily occurred when background regions were incorrectly predicted as PPE objects, whereas false negatives were dominated by PPE objects that were misclassified as background, indicating that missed detections contributed more substantially to the overall detection errors than incorrect PPE class assignments. In contrast, direct confusion between complementary classes was relatively limited. For example, only 7 No Helmet instances were predicted as Helmet On, while 1 No Vest instance was classified as Vest On and 1 Vest On instance was classified as No Vest.

Similarly, confusion between Gloves On and No Glove was minimal (1 and 2 instances, respectively). These findings indicate that the evaluated models were generally able to distinguish visually similar PPE categories despite moderate photometric saturation variations. Furthermore, the confusion matrix reflects the influence of class imbalance. Majority classes, such as Helmet On (51 true positives) and Vest On (65 true positives), achieved substantially better detection performance than minority classes, including Gloves On (1 true positive), No Boots (0 true positives), and No Vest (1 true positive). This observation is consistent with the imbalanced class distribution of the dataset, where minority PPE classes contained substantially fewer annotated instances than the majority classes, thereby limiting the learning process for these categories. The limited number of annotated samples available for these minority classes likely reduced the models' ability to learn representative patterns, resulting in lower recall and a higher number of false negatives.



YOLOv8 Original Image YOLOv11 Original Image YOLOv12 Original Image

YOLOv8 Saturation +30% YOLOv11 Saturation +30% YOLOv12 Saturation +30%

YOLOv8 Saturation -30% YOLOv11 Saturation -30% YOLOv12 Saturation -30%

**Figure 4.** Evaluation Model With a Confusion Matrix

### 3.3. Visualization of Detection Results

As part of the analysis, the visualization results in the form of bounding boxes are displayed to illustrate the object detection by each model on the test images. This visualization facilitates a visual comparison of detection performance between different YOLO versions in Figure 5.

**Figure 5.** Example of Visual Detection Results (Bounding Box) from YOLOv8, YOLOv11, and YOLOv12 original image, saturation level 30%, and -30%

### 3.4. Model Comparison Analysis

This study evaluates three versions of the YOLO-based object detection model, namely YOLOv8, YOLOv11, and YOLOv12. The testing was conducted using standard metrics for object detection such as Precision, Recall, mAP50, and mAP50-95. Additionally, testing was also conducted by altering the image saturation levels to determine the model's robustness against changes in visual conditions. YOLOv8 shows the most stable performance among the three models. This model achieved the highest scores on almost all evaluation metrics, both under original image conditions and when saturation was increased or decreased. This stability indicates that YOLOv8 has a good ability to detect various types of objects consistently. Meanwhile, YOLOv11 has a performance that falls between YOLOv8 and YOLOv12 but tends not to show dominance in certain aspects. In other words, YOLOv11 does not provide significant improvements compared to the previous version, both in terms of stability and the ability to detect small objects. On the other hand, YOLOv12, although not superior overall, shows significant improvement in the ability to detect small objects. This may be due to architectural changes, such as an increase in feature map resolution.

However, the overall performance of YOLOv12 remains lower than YOLOv8, indicating that improvements in specific object categories do not necessarily translate into better overall detection performance. The relatively lower mAP values obtained in this study compared to several previous PPE detection studies can be attributed to the complexity and imbalance of the dataset. Although the dataset reached 3,116 images after augmentation, the original dataset consisted of only 1,072 images. Furthermore, the detection task involved eight PPE classes with highly uneven object distributions. For example, the No Vest and Gloves On classes contained only 5 and 9 annotated objects, respectively, while the No Boots class contained no annotated samples. Such severe class imbalance may limit the model's ability to learn representative features for minority classes, thereby reducing overall detection performance. In addition, unlike many previous studies that focused on one or two PPE categories under standard imaging conditions, this study simultaneously evaluated eight PPE-related classes under photometric saturation variations (-30%, 0%, and +30%), resulting in a substantially more challenging detection problem.

Overall, the experimental results demonstrate that YOLOv8 provides the best balance between detection accuracy, robustness, and model stability under varying saturation conditions. Therefore, YOLOv8 can be considered the most suitable model for PPE detection applications in environments where image quality may vary due to changes in lighting and visual conditions. These findings also highlight the importance of robustness evaluation in addition to accuracy assessment when selecting object detection models for real-world occupational safety monitoring systems.

## 4. CONCLUSIONS

This study systematically evaluated the robustness of YOLOv8, YOLOv11, and YOLOv12 for Personal Protective Equipment (PPE) detection under controlled photometric saturation variations. The experimental results demonstrated that YOLOv8 achieved the best overall performance, with an average score of 0.495, outperforming YOLOv11 (0.490) and YOLOv12 (0.430). Although YOLOv11 produced competitive results and YOLOv12 showed relatively better performance for several small-object classes, YOLOv8 consistently provided the highest detection accuracy and stability across all saturation levels. Furthermore, the relatively small performance differences observed under saturation variations of −30%, 0%, and +30% indicate that all evaluated models maintained stable detection capabilities under moderate changes in image color characteristics.

Beyond the comparative evaluation of three recent YOLO architectures, this study contributes a systematic robustness evaluation framework based on controlled photometric saturation variations. The proposed evaluation provides additional evidence regarding model stability under realistic visual disturbances that commonly occur in industrial surveillance systems, such as variations in illumination, camera characteristics, and automatic color adjustment. From a practical perspective, the findings provide useful guidance for selecting robust object detection models for real-time occupational safety monitoring systems deployed in construction and industrial environments where imaging conditions cannot always be controlled.

Several limitations should be acknowledged. First, the comparative evaluation was conducted using a single deterministic training and testing configuration; therefore, repeated experiments with different random seeds were unavailable, preventing statistical significance analysis. Second, although photometric saturation augmentation was employed to simulate controlled color variations, the augmented images were generated from the same original dataset and may not fully represent the diversity of real industrial environments. Consequently, the generalizability of the evaluated models should be interpreted with appropriate caution.

Future research should validate the proposed robustness evaluation using larger and more diverse real-world PPE datasets with more balanced class distributions, particularly for minority classes such as Gloves On and No Vest. Repeated training using different random seeds is also recommended to enable statistical significance testing, such as the Friedman or Wilcoxon tests, thereby providing stronger evidence for comparative performance analysis. In addition, integrating robustness evaluation with Explainable Artificial Intelligence (XAI) techniques would improve model transparency and support the reliable deployment of computer vision systems for occupational safety applications.

## REFERENCES

Ahmed, M. I. B., Saraireh, L., Rahman, A., Al-Qarawi, S., Mhran, A., Al-Jalaoud, J., Al-Mudaifer, D., Al-Haidar, F., AlKhulaifi, D., Youldash, M., & Gollapalli, M. (2023). Personal protective equipment detection: A deep-learning-based sustainable approach. *Sustainability, 15*(18), 13990. https://doi.org/10.3390/su151813990

Biswas, S., Bhowmick, S., Islam, T., Ahmed, M. A., Islam, T., & Siddique, S. (2023). Real-time construction safety gear detection using YOLOv4 with Darknet. *Proceedings of the 17th IEEE International Conference on Applied Information and Communication Technology (AICT 2023),* 1–6. https://doi.org/10.1109/AICT59525.2023.10313207

Cahyani, S., Sari, A. K., & Harjoko, A. (2024). Underwater quality enhancement based on mixture contrast limited adaptive histogram and multiscale fusion. *International Journal of Advanced Computer Science and Applications, 15* (7), 647–655. https://doi.org/10.14569/IJACSA.2024.0150763

Cai, Z., Zhou, K., & Liao, Z. (2025). A systematic review of YOLO-based object detection in medical imaging: Advances, challenges, and future directions. *Computers, Materials & Continua, 85* (2), 2255–2303. https://doi.org/10.32604/cmc.2025.067994

De, K., & Pedersen, M. (2021). Impact of colour on robustness of deep neural networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW),* 21–30

Elesawy, A., Abdelkader, E. M., & Osman, H. (2024). A detailed comparative analysis of You Only Look Once-based architectures for the detection of personal protective equipment on construction sites. *Eng, 5*(1), 347–366. https://doi.org/10.3390/eng5010019

Gholinavaz, S., Saeedi, N., & Samadi Gharehveran, S. (2025). Robustness analysis of YOLO and Faster R-CNN for object detection in realistic weather scenarios with noise augmentation. *Scientific Reports, 15,* 44888. https://doi.org/10.1038/s41598-025-28737-5

Gyawali, S., Mohammadjafari, A., Ghimire, S., & Habibnezhad, M. (2026). Reliable vision-based PPE detection for construction safety in adverse environmental conditions. *Buildings, 16*(12), 2447. https://doi.org/10.3390/buildings16122447

Irawan, B., Andono, P. N., & Basuki, R. S. (2024). Optimization of YOLOv5 hyperparameter using Adam optimizer in vehicle object detection. *Journal of Applied Intelligent Systems, 9* (1), 40–50. https://doi.org/10.62411/jais.v9i1.9244

Jocher, G., Chaurasia, A., Qiu, J., & Ultralytics. (2023). *YOLOv8 documentation*. https://docs.ultralytics.com

Khanam, R., & Hussain, M. (2025). A review of YOLOv12: Attention-based enhancements vs. previous versions. *arXiv preprint arXiv:2504.11995*. https://arxiv.org/abs/2504.11995

Kim, D., & Xiong, S. (2025). Enhancing worker safety: Real-time automated detection of personal protective equipment to prevent falls from heights at construction sites using improved YOLOv8 and edge devices. *Journal of Construction Engineering and Management, 151* (1). https://doi.org/10.1061/JCEMD4.COENG-14985

Kisaezehra, Farooq, M. U., Bhutto, M. A., & Kazi, A. K. (2023). Real-time safety helmet detection using YOLOv5 at construction sites. *Intelligent Automation & Soft Computing, 36* (1), 911–927. https://doi.org/10.32604/iasc.2023.031359

Laily, M. E., Fajri, F. N., & Pratamasunu, G. Q. O. (2022). Deteksi penggunaan alat pelindung diri (APD) untuk keselamatan dan kesehatan kerja menggunakan metode Mask Region Convolutional Neural Network (Mask R-CNN). *Jurnal Komputer Terapan, 8* (2), 279–288. https://doi.org/10.35143/jkt.v8i2.5732

Li, Y., Zhang, J., Hu, Y., Zhao, Y., & Cao, Y. (2022). Real-time safety helmet-wearing detection based on improved YOLOv5. *Computer Systems Science and Engineering, 43* (3), 1219–1230. https://doi.org/10.32604/csse.2022.028224

Mailoa, R. M., & Santoso, L. W. (2022). Deteksi rompi dan helm keselamatan menggunakan metode YOLO dan CNN. Jurnal Infra, 10(2), 49–55.

Mair, Z. R., Harjoko, A., Gustriansyah, R., Cahyani, S., Heriansyah, R., Permatasari, I., & Irfani, M. H. (2025). An enhanced deep learning framework for diabetic retinopathy classification using multiple convolutional neural network architectures. *International Journal of Advanced Computer Science and Applications, 16* (11), 769–776. https://doi.org/10.14569/IJACSA.2025.0161176

Nazli, N. A. N. M., Sabri, N., Aminuddin, R., Ibrahim, S., Yusof, S., & Nasir, S. D. N. M. (2024). A real-time system for detecting personal protective equipment compliance using deep learning model YOLOv5. *Procedia Computer Science, 245*, 647–656. https://doi.org/10.1016/j.procs.2024.10.291

Ofori-Oduro, M., & Amer, M. (2024). Defending object detection models against image distortions. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024),* 3854–3863.

Ramos, L., & Sappa, A. D. (2025). A decade of You Only Look Once (YOLO) for object detection: A review. *IEEE Access, 13*, 192747–192794. https://doi.org/10.1109/ACCESS.2025.3630988

Roboflow. (2025). *Roboflow: Computer vision tools for developers and enterprises*. https://roboflow.com

Ruaz, F. M. (2023). Recognition of safety helmet wearing of operator local ground floor unit 2 Suralaya PGU based on improved YOLOv3. *Journal of Mechanical Design and Testing, 5* (1), 35–43. https://doi.org/10.22146/jmdt.85645

Santos, W., Lorente, A., Rojas, C., Mariscal, G., & Lorente, R. (2025). Impact of personal protective equipment in preventing occupational injuries: A systematic review and meta-analysis. *Frontiers in Public Health, 13*, 1720363. https://doi.org/10.3389/fpubh.2025.1720363

Sivanraj, S., Uduwage, D. N. L. S., & Tripathi, M. (2025). Comparison of YOLO algorithms for PPE compliance monitoring at construction sites. In K. G. A. S. Waidyasekara, H. S. Jayasena, P. L. I. Wimalaratne, & G. A. Tennakoon (Eds.), *Proceedings of the 13th World Construction Symposium* (pp. 438–450). https://doi.org/10.31705/WCS.2025.33

Wang, Z., Cai, Z., & Wu, Y. (2023). An improved YOLOX approach for low-light and small object detection: PPE on tunnel construction sites. *Journal of Computational Design and Engineering, 10* (3), 1158–1175. https://doi.org/10.1093/jcde/qwad042

Xiong, R., & Tang, P. (2021). Pose guided anchoring for detecting proper use of personal protective equipment. Automation *in Construction, 130*, 103828. https://doi.org/10.1016/j.autcon.2021.103828