



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Deep Learning for Personal Protection Equipment (PPE) Detection in Real-Life Scenarios

João Pedro Tomás Ferreira

Master's in Data Science

Supervisor:

PhD Tomás Gomes Silva Serpa Brandão, Assistant Professor,
ISCTE-IUL

Supervisor:

PhD Luís Miguel Martins Nunes, Associate Professor,
ISCTE-IUL

September, 2024

Department of Quantitative Methods for Management and Economics

Department of Information Science and Technology

**Deep Learning for Personal Protection Equipment (PPE)
Detection in Real-Life Scenarios**

João Pedro Tomás Ferreira

Master's in Data Science

Supervisor:
PhD Tomás Gomes Silva Serpa Brandão, Assistant Professor,
ISCTE-IUL

Supervisor:
PhD Luís Miguel Martins Nunes, Associate Professor,
ISCTE-IUL

September, 2024

*“Põe quanto és
No mínimo que fazes.”
Ricardo Reis*

Acknowledgment

I would like to extend my deepest gratitude to everyone who, in one way or another, has been part of this journey, which now comes to an end.

To my supervisors, Professor Tomás and Professor Luís, thank you for the invaluable knowledge you have shared with me, for your patience, and for your willingness to be part of this journey. I am deeply grateful to both of you.

To my colleagues from this Master's program, thank you for making this journey more enjoyable despite the challenges we faced together.

To my friends, thank you for your devoted support, not only during the development of this dissertation but also during the times I needed you the most. You are my second family, and I am proud to be part of it.

To my family, thank you for all the support you have given me when I needed it most. We may not be large in number, but we have remained strong and united through all that we have faced together.

To my mother and brother, the two persons that keep reminding me that everything is worth fighting for. Thank you for your patience, your support, and for helping me feel that I am on the right track and doing the right thing.

To my girlfriend Sara, the person that I love the most. Thank you for always believing in me, for all the support and patience, and for appearing in my life. Words cannot describe how thankful I am for you. I am really proud of what we been achieving together.

Lastly, to my father, my *estrelinha*. There isn't a day when I don't think of you. As time goes by, I miss you more and more, and how much better this phase would have been with you by my side. Still, I hope you're proud of what I've become and what I've achieved in the same way that I'm proud to be your son.

Resumo

A detecção de Equipamentos de Proteção Individual (EPI) desempenha um papel fulcral na garantia do bem-estar dos trabalhadores, prevenindo lesões graves e fatalidades (SIF). Esta dissertação centra-se no desenvolvimento e implementação de modelos de Aprendizagem Profunda (AP) para a monitorização da utilização de EPI, com particular ênfase na eficácia do modelo de detecção de objectos YOLOv8.

Foi realizada uma revisão da literatura que destacou o YOLOv8 como a abordagem mais promissora devido ao seu equilíbrio entre precisão, velocidade e adaptabilidade em tarefas de detecção de objectos. Esta visão guiou o processo de desenvolvimento, onde os modelos YOLOv8 de médio e grande porte foram treinados e avaliados em imagens do mundo real para detetar e classificar diferentes tipos de EPI. Os modelos alcançaram aproximadamente 80% de precisão e 70% de revocação, demonstrando a sua eficácia na identificação e categorização de EPI, tornando-os candidatos adequados para aplicações práticas em vários ambientes industriais.

Com base nisto, foram desenvolvidas várias estratégias de monitorização, aproveitando o YOLOv8 para o acompanhamento contínuo da conformidade do EPI em ambientes dinâmicos. Apesar de desafios como a variabilidade ambiental, os sistemas revelaram-se eficazes em testes práticos, demonstrando o seu potencial para implementação em operações industriais críticas para a segurança.

Este trabalho também explorou as implicações práticas da implementação de tais sistemas, revelando melhorias significativas na segurança do local de trabalho. Ao automatizar o processo de monitorização, o sistema reduz o erro humano, assegura a adesão aos protocolos de segurança e melhora a gestão global do cumprimento dos EPI em ambientes de trabalho perigosos.

Abstract

The detection of Personal Protective Equipment (PPE) plays a pivotal role in ensuring the well-being of workers, preventing serious injuries and fatalities (SIF). This dissertation focuses on the development and implementation of Deep Learning (DL) models for monitoring of PPE usage, with particular emphasis on the effectiveness of the YOLOv8 object detection model.

A literature review was conducted which highlighted YOLOv8 as the most promising approach due to its balance of accuracy, speed, and adaptability in object detection tasks. This insight guided the development process, where the YOLOv8 medium and large models were trained and evaluated on real-world images to detect and classify different types of PPE. The models achieved approximately 80% precision and 70% recall, demonstrating their effectiveness in identifying and categorising PPE, making them suitable candidates for practical applications in various industrial environments.

Building on this, three monitoring processes were developed, leveraging YOLOv8 for continuous PPE compliance tracking in dynamic settings. Despite challenges such as environmental variability, those processes proved effective in practical tests, demonstrating their potential for deployment in safety-critical industrial operations.

The research also explored the practical implications of implementing such processes, revealing significant improvements in workplace safety. By automating the monitoring process, the processes reduce human error, ensure adherence to safety protocols, and enhance the overall management of PPE compliance in hazardous work environments.

Contents

Acknowledgment	iii
Resumo	v
Abstract	vii
List of Tables	xi
List of Figures	xiii
Glossary of Abbreviations and Acronyms	xv
Chapter 1. Introduction	1
1.1. Motivation	1
1.2. Research Questions	2
1.3. Objectives	2
1.4. Research Methodology	3
1.5. Dissertation Structure	4
Chapter 2. Background Concepts and Related Work	7
2.1. Personal Protective Equipment (PPE)	7
2.2. Literature Review Methodology	8
2.3. Related Work	9
2.4. Summary	12
Chapter 3. Proposed PPE Detection Processes	15
3.1. YOLOv8 architecture	15
3.2. Proposed Processes architecture	18
3.2.1. Unified Detection Process (UDP)	18
3.2.2. Parallel Detection Process (PDP)	19
3.2.3. Sequential Detection Process (SDP)	20
3.3. Data Preparation and Assessment Metrics	21
3.3.1. Training and Validation Data	21
3.3.2. Testing Data	22
3.3.3. Assessment metrics	23
Chapter 4. Training of Models	27
4.1. All Classes Detection (Person & PPE)	27
4.2. Person detection	28

4.3. PPEs detection	32
Chapter 5. Results and Discussion	39
5.1. Unified Detection Process (UDP)	39
5.2. Parallel Detection Process (PDP)	41
5.3. Sequential Detection Process (SDP)	43
5.4. Comparative Analysis with Previous Studies	45
Chapter 6. Conclusion	47
6.1. Final Remarks	47
6.2. Future Work	48
Bibliography	51
Appendix A. Submitted Article	55

List of Tables

1	YOLOv8 Model Variants Performance Comparison (adapted from [1])	18
2	Results of YOLOv8 Model Training for Detection of All Classes (Results expressed in percentage)	27
3	Results of YOLOv8 Models Trained for Person Detection (Results expressed in percentage)	30
4	Results of YOLOv8 Models Validated for Person Detection (Results expressed in percentage)	30
5	Results of YOLOv8 Models Training for PPE Detection. (Results expressed in percentage)	34
6	Performance of Unified Detection Process for each test dataset (Results expressed in percentage)	39
7	Performance of Parallel Detection Process for each test dataset (Results expressed in percentage)	42
8	Performance of Sequential Detection Process for each test dataset (Results expressed in percentage)	43

List of Figures

1	CRISP-DM methodology steps (from [2])	4
2	Example of different types of PPE (Adapted from [3])	8
3	Distribution of articles by year of publication	9
4	YOLOv8 model architecture (from [4])	16
5	Workflow of UDP from input image to output with detected objects	19
6	Workflow of PDP from input image to output with detected persons and PPE items	20
7	Overview of SDP: A two-stage detection process with image cropping for enhanced precision	21
8	IoU calculation	24
9	Loss and metrics curves during the training of All_classes model	28
10	Loss and metrics curves during the training of Person model using YOLOv8S	29
11	Loss and metrics curves during the training of Person model using YOLOv8M	30
12	Comparison of person detection performance between the default YOLOv8 and the trained YOLOv8 model.	31
13	Example of image used in the training of the models	32
14	Example of a cropped image used in the training of the models	33
15	Example of a masked image used in the training of the models	33
16	Loss and metrics curves during the training of PPE_all model using YOLOv8M	35
17	Loss and metrics curves during the training of Cropped_PPE_all model using YOLOv8M	36
18	Loss and metrics curves during the training of Masked_PPE_all model using YOLOv8M	37
19	Confusion Matrix for UDP on Test Data 1	40
20	Confusion Matrix for UDP on Test Data 2	41
21	Confusion Matrix for PDP on Test Data 1	42
22	Confusion Matrix for PDP on Test Data 2	43
23	Confusion Matrix for SDP on Test Data 1	44

Glossary of Abbreviations and Acronyms

ANN - Artificial Neural Networks
APAS - Average Precision Across Scales
BLE - Bluetooth Low Energy
CNNs - Convolutional Neural Networks
CRISP-DM - Cross Industry Standard Process for Data Mining
DL - Deep Learning
EU - European Union
FN - False Negative
FP - False Positive
IoT - Internet of Things
IoU - Intersection over Union
mAP - Mean Average Precision
ML - Machine Learning
NAS-FPN - Neural Architecture Search-Feature Pyramid Network
OAN - Operator Area Networks
PPE - Personal Protection Equipment
R-CNN - Region-based Convolutional Neural Network
RFID - Radio-frequency Identification
SIF - Serious Injuries and Fatalities
SSD - Single Shot MultiBox Detector
TP - True Positive
YOLO - You Only Look Once

CHAPTER 1

Introduction

The proper use of Personal Protection Equipment (PPE) is crucial in preventing Serious Injuries and Fatalities (SIF), significantly reducing the hazards faced by workers. According to Eurostat data reported in [5] related to the year of 2021, there were 2.88 million non-fatal work-related accidents in the European Union (EU) that resulted in at least four days of absence from work. Additionally, around 3,300 of these incidents were fatal, emphasising the importance of effective PPE utilisation to ensure worker safety.

The study presented in [6] highlighted common issues such as using inappropriate PPE, not wearing PPE at all, or incorrect usage. These errors can lead to severe consequences, including cuts, burns, respiratory problems, and even death. To address this widespread problem, it is essential for employers to prioritise thorough training on the selection, usage, and maintenance of PPE, and also monitor the correct use of such equipments.

Despite the availability of PPE, its effectiveness is often compromised by human factors such as lack of awareness, improper fit, and discomfort, leading to inconsistent usage, as explored in [7]. Additionally, the rapid development of industrial environments and technologies presents new risks and challenges that traditional PPE may not adequately address.

Therefore, continuous advancements in PPE technology and monitoring systems are imperative. The integration of PPE monitoring systems within worksites to verify the correct use of PPE and alert when it is not used properly offers a promising solution to enhance safety protocols.

This dissertation explores the development of various Deep Learning (DL) models trained to detect and classify different PPEs using several Computer Vision approaches, and their implementation in different processes to monitor the correct use of PPE.

1.1. Motivation

The motivation behind this research stems from the critical need to enhance workplace safety and reduce the incidence of SIF among workers. Despite the widespread availability of PPE, improper use, non-compliance, and inadequate monitoring remain significant challenges. Traditional methods of PPE monitoring rely heavily on manual inspection and worker compliance, which are often inconsistent and prone to human error. This inadequacy assures the necessity for automated and reliable systems that can secure the proper use of PPE in real-time.

Advancements in Computer Vision and DL technologies present an opportunity to develop complex models capable of accurately detecting and classifying PPE. These technologies can transform the way PPE compliance is monitored, shifting from manual checks to automated systems that provide real-time feedback and alerts.

The potential impact of integrating Computer Vision and DL into PPE monitoring systems is substantial. By automating the detection and classification of PPE, these systems can provide continuous surveillance, identify non-compliance instantly, and generate alerts to prevent hazardous situations. This, not only enhances worker safety, but also promotes a culture of safety within companies, leading to better adherence to safety protocols and potentially saving lives.

Furthermore, the rapid development of industrial environments and the introduction of new technologies continuously introduce new risks and challenges. Traditional PPE and monitoring methods may not be sufficient to address these evolving hazards. Therefore, it is necessary to leverage new technologies to develop advanced monitoring systems that can adapt to and mitigate these risks effectively.

1.2. Research Questions

The primary aim of this dissertation is to investigate and address key questions related to the development and application of DL models for monitoring the use of PPE. To achieve this objective, the following research questions have been defined:

- **RQ1:** How effective are YOLOv8 models in detecting and classifying PPE in real-world scenarios?
- **RQ2:** How do processes based on DL models perform in PPE monitoring, and what are the challenges associated with their implementation?
- **RQ3:** Is it beneficial to combine person detection and PPE detection to build a more effective monitoring process?

1.3. Objectives

Taking into consideration the Motivation and Research Questions outlined in the previous sections, the main objective of this dissertation is to develop and evaluate processes that utilise various DL models for monitoring the use of PPE through image-based training.

First, the dissertation focuses on using different object detectors, including the latest YOLOv8 model, to detect and classify various types of PPE in real-world images following distinct approaches. This involves evaluating the effectiveness of these models in accurately identifying and categorising PPE.

Second, the dissertation addresses the design and development of integrated processes that leverage these DL models for PPE monitoring. This includes examining how well these processes perform in practical settings and identifying any challenges or limitations they may encounter.

Additionally, the research explores the practical benefits of combining person detection and PPE detection. The objective is to determine if a monitoring process is more effective when these two detection tasks are integrated.

Finally, a comprehensive evaluation of the developed processes is conducted, assessing their strengths and weaknesses. The dissertation also proposes improvements based on the findings and explores potential future advancements in DL and Computer Vision technologies for PPE monitoring.

1.4. Research Methodology

This work was conducted based on the Cross Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a widely recognised methodology for data mining that offers a structured approach to planning and executing data analysis projects. It consists of six phases, as illustrated in Figure 1, including Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment [8]. Each phase is detailed as follows:

- **Business Understanding:** This phase involves defining the project objectives and requirements from a business perspective. For this dissertation, it focused on understanding the need for enhanced PPE monitoring systems and setting goals for implementing DL models.
- **Data Understanding:** Initial data is collected and explored to gain insights and identify any data quality issues. This research involved gathering images of PPE and analysing their characteristics to ensure they were suitable for model training.
- **Data Preparation:** This phase covers data cleaning, transformation, and preparation for modelling. In this work, data preparation included annotating images with bounding boxes, splitting the dataset into training, validation, and test subsets, and adjusting images according to the specific modelling approaches employed.
- **Modelling:** This phase involves selecting and applying various modelling techniques to the prepared data. In this dissertation, different YOLOv8 models were trained to detect and classify PPE in images, with each model representing a distinct approach.
- **Evaluation:** The models are assessed based on their performance, and results are compared to the initial objectives. This phase focused on evaluating the accuracy and effectiveness of the models in real-world scenarios.
- **Deployment:** The final phase involves synthesising the findings, methodologies, and results into a comprehensive document. In this case, it focused on compiling the dissertation, detailing the research process, and presenting conclusions and recommendations based on the study.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives Background Business Objectives Business Success Criteria Assess Situation Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits Determine Data Mining Goals Data Mining Goals Data Mining Success Criteria Produce Project Plan Project Plan Initial Assessment of Tools and Techniques	Collect Initial Data Initial Data Collection Report Describe Data Data Description Report Explore Data Data Exploration Report Verify Data Quality Data Quality Report	Data Set Data Set Description Select Data Rationale for Inclusion / Exclusion Clean Data Data Cleaning Report Construct Data Derived Attributes Generated Records Integrate Data Merged Data Format Data Reformatted Data	Select Modeling Technique Modeling Technique Modeling Assumptions Generate Test Design Test Design Build Model Parameter Settings Models Model Description Assess Model Model Assessment Revised Parameter Settings	Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models Review Process Review of Process Determine Next Steps List of Possible Actions Decision	Plan Deployment Deployment Plan Plan Monitoring and Maintenance Monitoring and Maintenance Plan Produce Final Report Final Report Final Presentation Review Project Experience Documentation

FIGURE 1. CRISP-DM methodology steps (from [2])

1.5. Dissertation Structure

This dissertation is structured into six chapters, each addressing a critical aspect of the research and development process. The structure is designed to guide through the context and motivations of the study, the state-of-the-art in PPE detection and monitoring processes, the methodologies employed, the results obtained, and the conclusions that were made. This dissertation will be presented in the following chapters:

- **Chapter 1** introduces the dissertation, outlining the context, motivations, objectives, contributions, research questions, and methodology. It establishes the foundation for the study and explains the significance of the research in the field of PPE monitoring.
- **Chapter 2** explores the current state-of-the-art in PPE detection and monitoring. It reviews various types of PPE and summarises previous research in the area. The chapter highlights the strengths and weaknesses of different approaches, with the literature review conducted using the Systematic Literature Review methodology [9].
- **Chapter 3** details the methodologies employed in developing and implementing the PPE monitoring processes. This includes the YOLOv8 version, process architecture, data preparation phase, and evaluation metrics used to analyse the results.

- **Chapter 4** focuses on the training of the models and the specific approaches followed for each process. This chapter discusses the practical implementation of each approach and how it was tailored to achieve the desired performance.
- **Chapter 5** presents and discusses the results of the developed processes. It evaluates their performance and discusses their applicability in real-world scenarios, including practical implications and potential limitations. Additionally, this chapter includes a comparative analysis with previous studies, highlighting how the results obtained in this work align with or differ from those in the existing literature.
- **Chapter 6** summarises the key findings of the dissertation, reflects on the work conducted, and discusses potential improvements. It also outlines future directions for research and development in PPE monitoring systems.

CHAPTER 2

Background Concepts and Related Work

As workplaces evolve, the need for advanced safety measures becomes increasingly crucial. The landscape of workplace safety is undergoing a transformation with the adoption of new technologies. Real-time PPE detection systems, leveraging innovations like Bluetooth Low Energy (BLE), Radio-frequency Identification (RFID), and cutting-edge Computer Vision DL-based Approaches, are replacing conventional methods.

This chapter lays the groundwork for an in-depth exploration of both traditional and DL approaches to PPE detection. Before delving into the methodologies that were already used within this area of study, it is essential to understand the significance of PPE and its diverse types in ensuring a secure working environment. This chapter provides insights into the present and future of workplace safety regulations, emphasising the pivotal role of real-time PPE detection systems in safeguarding individuals in various occupational settings.

2.1. Personal Protective Equipment (PPE)

PPE refers to specialised gear or clothes designed to protect individuals from various workplace hazards, ensuring their safety and well-being. The types of PPE can vary based on the nature of the work environment and the specific risks involved. Figure 2 presents the most common types of PPE, such as:

- **Head Protection:** Helmets or hard hats safeguard against head injuries.
- **Eye and Face Protection:** Safety glasses, goggles, or face shields shield the eyes and face from debris, chemicals, or intense light.
- **Respiratory Protection:** Masks or respirators help filter out harmful airborne substances.
- **Hearing Protection:** Earplugs or earmuffs protect against excessive noise levels.
- **Hand and Arm Protection:** Gloves and arm coverings shield against cuts, burns, or chemical exposure.
- **Foot and Leg Protection:** Steel-toed boots or other specialised footwear safeguard against foot injuries.
- **Body Protection:** Coveralls, vests, or full-body suits provide overall protection against various hazards.

Understanding the importance of proper PPE usage is crucial for maintaining a safe working environment. Real-time PPE detection systems play a vital role in ensuring compliance and addressing potential risks promptly.



FIGURE 2. Example of different types of PPE (Adapted from [3])

2.2. Literature Review Methodology

A systematic literature review was conducted between October and November 2023, and later revised in June 2024. As a starting point, it was utilised the databases of IEEE, Elsevier, and articles published in Multimedia Tools and Applications. The objective was to comprehensively explore the state of the art in real-time PPE detection. The keywords considered for this search were “Personal Protection Equipment (PPE) detection” and “Computer Vision”. Furthermore, manual searches of references in the identified articles were also conducted to ensure the inclusion of relevant studies and to uncover any additional sources that might contribute valuable insights. This process led to the discovery of a significant study published in 2013.

The inclusion criteria for the review focused on studies that addressed methodologies for real-time PPE detection, with an emphasis on technologies such as Traditional Approaches and DL Approaches. Additionally, only studies contributing with insights into workplace safety and published between 2019 and 2024 were included.

The search process retrieved a total of 30 articles directly related to the topic. This systematic approach offers a robust foundation for the subsequent analysis of both traditional and DL approaches to real-time PPE detection. Figure 3 shows the distribution of articles by year.

Regarding each article considered in this review, detailed information was extracted, including its title, authors, year of publication, methodology/approach, the dataset used, and a brief summary of the conducted work in the study.

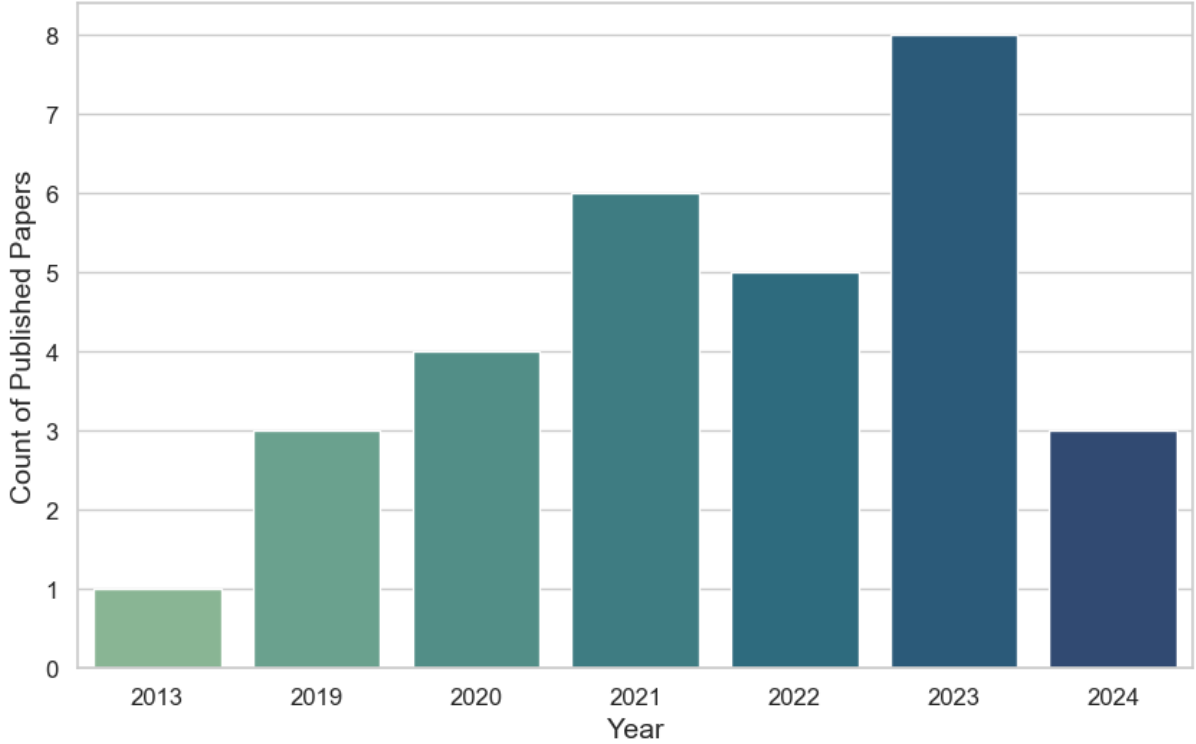


FIGURE 3. Distribution of articles by year of publication

For the subsequent data analysis, a narrative synthesis approach was carried out. This method involved systematically summarising the findings of each study and identifying the key methodologies that were implemented. Through this analysis, the current state of the art in real-time PPE detection was perceived. This not only facilitated the definition of future work within the scope of this study, but also aimed to provide other researchers with valuable insights, guiding them in determining their possible approaches to this workplace safety field.

2.3. Related Work

Traditional methods for enhancing workplace safety have leveraged established technologies to tackle the challenges associated with real-time PPE detection. These approaches, often built on mature technologies such as Bluetooth Low Energy (BLE), Radio-frequency Identification (RFID), Operator Area Networks (OANs), and the Internet of Things (IoT), have been widely adopted to address PPE compliance monitoring in various industries.

BLE has been used to mitigate the limitations of manual and visual PPE monitoring, which are labour-intensive and error-prone. For instance, in [10], it was introduced a real-time PPE detection system using BLE tags and receivers to prevent the use of power tools without proper PPE, thereby reducing injury risks.

RFID technology has also been employed to automate PPE compliance monitoring. Authors of [11] developed a mobile RFID portal for construction sites, combining RFID tags, readers, and a user interface to provide real-time feedback and extensive data for safety analysis.

OANs offer localised communication networks between workers and their PPE. In [12] it proposed an OAN-based system using wearable tags and a centralised server to enhance real-time PPE compliance monitoring and generate data for targeted safety interventions.

The arrival of Smart IoT technologies provides another solution for automating PPE-tool pair checks in construction settings. In [13] it was proposed a system integrating sensors, wireless networks, and a central monitoring system to offer real-time monitoring and data-driven insights for improved safety.

While these traditional approaches offer significant improvements over manual methods of PPE monitoring, they each have limitations in terms of scalability, accuracy, and adaptability to diverse environments. BLE and RFID systems, for instance, can face challenges with interference and require infrastructure that may not always be feasible to deploy in certain settings. OANs and IoT solutions, on the other hand, are more flexible but often involve higher costs and more complex implementation procedures.

In contrast to these traditional methods, the emergence of Computer Vision and DL-based approaches represents a powerful alternative for real-time PPE detection, offering enhanced accuracy, adaptability, and scalability. Object detection algorithms, underpinned by DL models such as Convolutional Neural Networks (CNNs), have significantly advanced, enabling machines to perform PPE detection with high precision. DL models, particularly CNNs, have become the backbone of modern object detection algorithms [14].

Object detection algorithms can be broadly categorised into two main types: two-stage and one-stage detectors [15, 16]. Two-stage detectors, such as R-CNN and Faster R-CNN, first generate object region proposals and then classify each proposal and refine its bounding box [14]. In contrast, one-stage detectors, such as Single Shot MultiBox Detector (SSD) and YOLO, directly predict bounding boxes and class probabilities for objects in a single stage, reducing computational cost [16, 15].

In the context of PPE detection, several studies have explored the use of two-stage object detection architectures like Faster R-CNN. For instance, in [17] it was utilised a Faster R-CNN architecture to develop a system that achieved approximately 80% accuracy and precision in detecting workers not fully compliant with PPE requirements. However, this approach primarily focused on detecting PPE rather than the individuals wearing it, which limits its applicability in scenarios where accurate detection of both the person and their PPE is crucial. Additionally, in [18] was leveraged a Faster R-CNN model trained on the CHVG dataset [19], which includes various PPE classes such as helmets, vests, and safety glasses.

SSD-based architectures have demonstrated significant efficiency for real-time PPE detection. Authors of [20] proposed a system using reverse progressive attention to combine features from multiple scales into an SSD detector, achieving a mAP of 83.9%. Similarly, [21] developed an SSD-based method using MobileNet to detect workers without hardhats at construction sites, with precisions of 87.4% for non-compliant and 89.4% for

compliant individuals. Additionally, [22] applied an SSD architecture with VGG-16 in the electrical sector, prioritising speed and accuracy, and demonstrating better performance compared to YOLO and Faster R-CNN for PPE detection. It is also worth mention that in [23] it was developed a hard-hat monitoring system that combined SqueezeNet for object detection and MobileNet for image classification.

YOLO models have become popular in PPE detection due to their balance of speed and accuracy, making them well-suited for real-time monitoring across various sectors such as construction, industrial safety, and healthcare.

Starting with YOLOv3, [24] develops a system to classify workers as either "compliant" or "non-compliant" with PPE requirements. Achieving an impressive mean Average Precision (mAP) of 93%, the system also included a user-friendly GUI for practical application. Similarly, authors of [25] used YOLOv3 to create a real-time model for monitoring PPE compliance on construction sites. Another study in [26] applied YOLOv3 to detect PPE and recognise faces in real-time, achieving up to 98% accuracy for six key types of PPE and 96% for face recognition. This fully automated system offered significant improvements over manual inspections, providing a reliable solution for ensuring worker safety in high-risk environments. In more specialised settings, such as offshore drilling platforms, it was used YOLOv3 with DL learning using ResNet50 to detect safety helmets and uniforms, achieving accuracies of 94.8% and 95.4%, respectively [27]. Additionally, it was explored in [28] the use of YOLOv3 to detect helmets in images combining it with face recognition.

Additionally, and moving to YOLOv4, [29] proposed a system that expanded PPE detection to include fire hazards, achieving a mAP of 79% across six different classes.

Other versions, like YOLOv5, have shown even greater efficiency. [30] combined YOLOv5 with an attention mechanism for enhanced PPE detection, while [31] compared several YOLOv5 models, with YOLOv5x achieving the highest mAP of 86.6% and YOLOv5s reaching the fastest processing speed at 52 frames per second. These studies highlight YOLOv5's ability to balance accuracy and speed, making it ideal for real-time safety applications.

Despite the significant progress in this domain, not all studies focus on detecting both the PPE and the individuals wearing it, which is crucial for effective workplace safety monitoring. Most recent research on YOLOv8, such as that conducted in [32] and [33], demonstrates the model's capability to detect various PPE items and the workers using them, achieving precisions of 80.6% and 86%, respectively. These studies illustrate the growing trend of integrating person detection with PPE monitoring, which is a key innovation explored in this dissertation. On the other hand, in [34], it was used YOLOv8 to detect helmets, gloves and shoes.

YOLO models have also been extensively used in healthcare settings during the COVID-19 pandemic to detect PPE such as masks and face shields. For instance, [35]

applied YOLOv3 for medical PPE detection, achieving a mAP of 96%, though varying accuracies across different PPE types presented challenges. Similarly, [36] and [37] used YOLOv4 for face mask detection, achieving accuracies of up to 99.3%, proving the importance of YOLO models in ensuring PPE compliance in healthcare environments. Additionally, in [38] and [39] it was used YOLOv5 to detect PPEs related to medical activities.

Another key application of DL is pose estimation, which allows machines to determine the positions and orientations of body parts from images and videos. Pose estimation can be categorised into single-pose and multi-pose methods. Single-pose algorithms track body joints for one person, while multi-pose methods handle multiple people, managing challenges like occlusions and varied poses [40].

Recent studies have integrated pose estimation with PPE detection to enhance real-time monitoring. For instance, in [41] it was developed a model combining a YOLOv3-based person detector, a Part Affinity Fields pose estimator, and a MobileNetv2 classifier with spatial attention. This approach achieved high precision, recall, and F1-scores (90% to 99%) for helmets, boots, and masks. Similarly, authors in [42] proposed a system using 2D keypoints from a pose estimator to localise and classify PPE, demonstrating high accuracies: 92% for hardhats, 98% for vests, 86% for gloves, and 95% for eyeglasses. On the other hand, in [43] it was introduced a pose-guided framework for multi-class PPE compliance detection, achieving high F1-scores for hardhats and safety vests. Additionally, in [44] it was employed the HigherHRNet pose estimator with MobileNetV2 for detecting PPE across 18 types and 5 body parts, achieving around 90% in accuracy, precision, recall, and F1-score, showcasing adaptability across various PPE types.

These advancements highlight the importance of integrating pose estimation into PPE detection systems. By accurately detecting both the PPE and the individuals wearing it, these methods address challenges such as misdetections and improve overall safety management. This integration aligns closely with the innovations in this dissertation, which focus on enhancing detection accuracy for both PPE and the users, and addressing issues related to erroneous detections.

2.4. Summary

This systematic literature review provides a comprehensive examination of advancements in real-time PPE detection. It highlights the current state of the art in both Traditional Approaches and DL-based Approaches.

The analysis of PPE monitoring techniques reveals two primary categories:

- **Traditional Approaches:** Technologies such as Bluetooth Low Energy (BLE), Radio-frequency Identification (RFID), Operator Area Networks (OAN), and Smart Internet of Things (IoT) have been explored in studies by [10], [11], [12], and [13], respectively. These methods offer high precision and require minimal changes to existing infrastructure but face challenges including performance issues in harsh environments and concerns over privacy.

- **Computer Vision Deep Learning-based Approaches:**

- **Object Detectors:** Algorithms such as YOLO, Faster R-CNN, and Single Shot MultiBox Detector (SSD) enable real-time processing with high accuracy. However, they encounter challenges such as variations in accuracy, training complexities, and limitations in detecting diverse types of PPE.
- **Pose Estimators:** Technologies explored by [41], [43], and [44] provide enhanced spatial understanding and accurate detection of body parts. Despite their advantages, these methods face issues related to model complexity, computational demands, and the difficulty in detecting specific types of PPE.

Key areas for future research include several promising directions aimed at advancing PPE compliance monitoring systems. Firstly, there is a need for the development of systems that can recognise a comprehensive range of PPE items. This would ensure consistency across various industries by standardising the classification of protective equipment.

Another crucial area is the integration of object detectors for person identification within compliance systems. By enhancing the precision of these systems, it becomes possible to more accurately identify individuals and ensure they are adhering to PPE requirements.

Additionally, innovating advanced person tracking techniques is essential for dynamic and effective monitoring of PPE usage over time. This advancement would provide more granular insights into PPE compliance and usage patterns.

The adoption of the latest versions of YOLO, such as YOLOv8 or even more recent ones such as YOLOv9 or YOLOv10, represents a significant opportunity for improving performance and leveraging updated features for robust PPE detection. Keeping pace with advancements in object detection algorithms can enhance the efficacy of compliance monitoring systems. In the case of this dissertation, it was used the YOLOv8 version since the YOLOv9 version only came out around February of 2024 and the YOLOv10 version came out during May 2024.

Finally, exploring diverse pose estimator architectures, such as OpenPose, DeepCut or even YOLOv8's pose estimator architecture, is important for improving the accuracy and spatial understanding of PPE compliance monitoring. Investigating new models and incorporating advanced neural network designs could lead to more precise detection and better overall system performance.

By focusing on these future research directions, the field of PPE compliance monitoring can advance towards more accurate, inclusive, and adaptable systems, enhancing overall workplace safety.

Proposed PPE Detection Processes

This section outlines the methodology employed to develop and evaluate three different processes for monitoring the use of PPE. Each process is designed with varying approaches to detect persons and their use of PPE, leveraging computer vision techniques and DL models. All models used in this study were trained using the 8th version of YOLO (YOLOv8), available at [1].

3.1. YOLOv8 architecture

YOLOv8, one of the major advancements in the YOLO series of object detection models, presented in [45], introduced significant improvements in both network architecture and performance. The structure of YOLOv8 is divided into two main components: the backbone network and the detection head.

As presented in [45], the backbone network utilises EfficientNet-B4, a state-of-the-art neural network architecture. EfficientNet-B4, a 71-layer network with 19 million parameters, processes input images resized to 512x512 pixels and outputs five multi-scale feature maps. This allows YOLOv8 to extract detailed features from images, enhancing its detection capabilities.

The detection head in YOLOv8 is based on Neural Architecture Search-Feature Pyramid Network (NAS-FPN), a neural architecture search method that creates feature pyramid networks for multi-scale predictions. YOLOv8 employs NAS-FPN-Cell, a sub-network that processes feature maps from the backbone network and generates bounding box predictions using anchor boxes to improve accuracy. Each feature map predicts three bounding boxes, resulting in a total of 15 bounding boxes for each input image [45]. Figure 4 presents the model’s architecture in detail.

YOLOv8 also introduces several new features and enhancements that set it apart from previous versions. It adopts Focal Loss, a loss function that focuses on difficult examples and reduces the impact of easy ones, thereby improving recall and precision. For data augmentation, YOLOv8 uses Mixup, which blends two images and their labels to create new training examples, increasing data diversity and reducing over-fitting. Additionally, YOLOv8 employs Average Precision Across Scales (APAS) as a new evaluation metric, measuring detection precision across different object scales and providing a more comprehensive performance assessment [45]. These innovations make YOLOv8 a robust and efficient model for object detection, addressing many limitations of earlier YOLO versions.

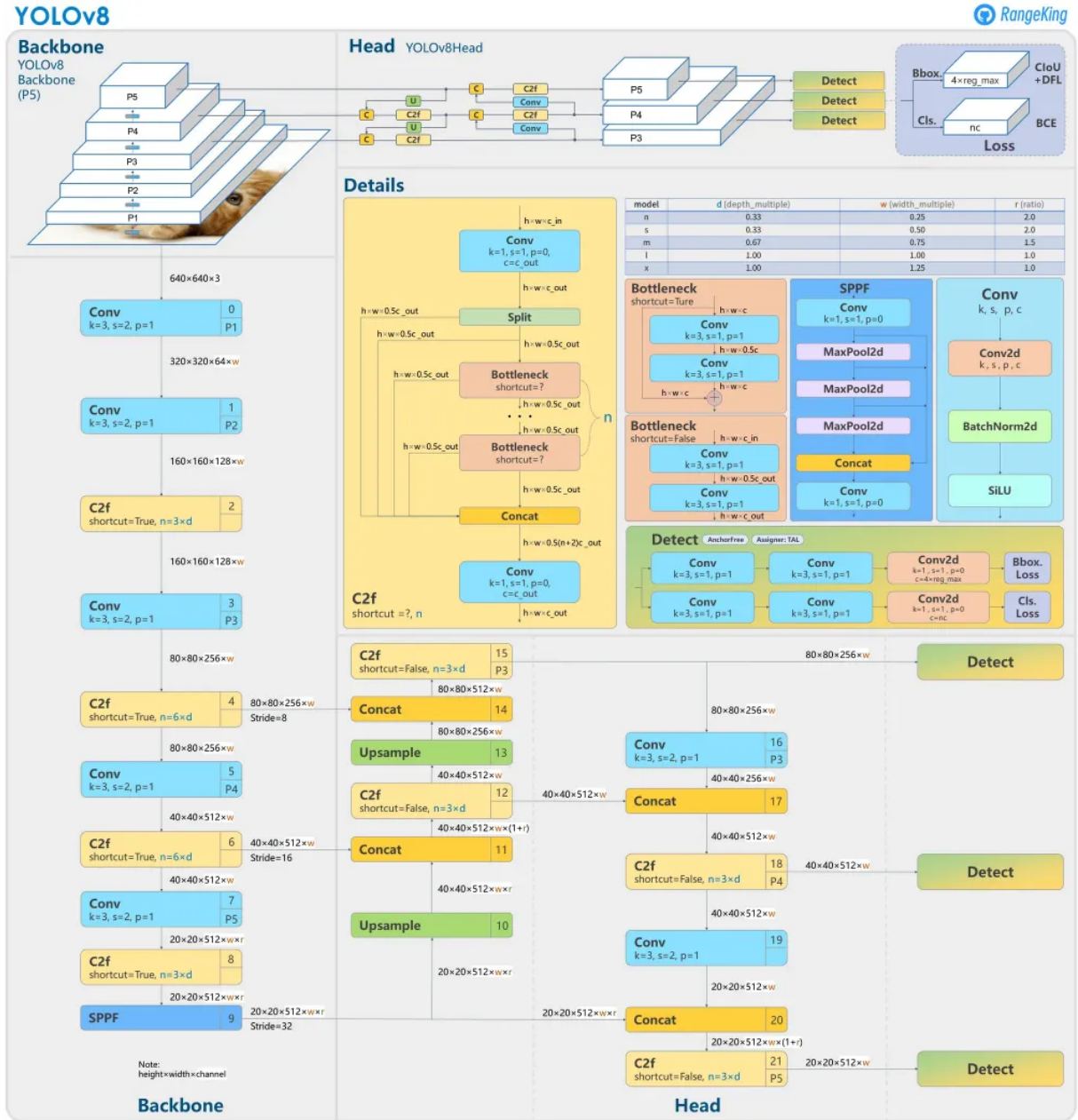


FIGURE 4. YOLOv8 model architecture (from [4])

Another innovation introduced in YOLOv8 is the model's ability to perform various tasks within the field of Computer Vision, such as image classification, instance segmentation, oriented bounding box object detection, and pose estimation, in addition to the object detection tasks handled by previous models.

YOLOv8 offers a range of model variants tailored to different computational and accuracy requirements, with each variant available in multiple sizes: *Nano*, *Small*, *Medium*, *Large*, and *Extra Large*. These variants are designed to accommodate various tasks within the field of computer vision, including object detection, image classification, instance segmentation, oriented bounding box detection, and pose estimation [1]. The *Nano* model is optimised for scenarios where computational resources are limited, providing faster inference times at the expense of some accuracy. On the other hand, the *Small* and *Medium*

models strike a balance between speed and precision, making them suitable for real-time applications with moderate resource availability. The *Large* and *Extra Large* models, while more computationally demanding, offer better detection performance and are ideal for tasks where precision is paramount. By offering this range of sizes, YOLOv8 ensures that users can select a model variant that best fits their specific application requirements, whether prioritising speed, accuracy, or resource efficiency.

As shown in Table 1, the YOLOv8 model variants differ in several key aspects, which are crucial for selecting the appropriate model based on specific application needs. Each column of the table provides important information about these aspects:

- **Size (pixels):** This column shows the input image resolution used for each model variant, which is set at 640x640 pixels across all YOLOv8 models. This standard size ensures a consistent comparison of performance metrics across the different variants.
- **mAP 50-95:** The mean Average Precision (mAP) is a widely used metric for evaluating the performance of object detection models. The mAP 50-95 column represents the average precision across a range of Intersection over Union (IoU) thresholds from 50% to 95%. A higher mAP score indicates better accuracy in detecting objects across different sizes and conditions. In the table, the *Extra Large* model (YOLOv8x) achieves the highest mAP score of 53.9, showing superior detection performance when compared to other variants.
- **Speed CPU ONNX (ms):** This column reflects the inference speed of each model variant when deployed on a CPU using the ONNX (Open Neural Network Exchange) format. The value is given in milliseconds (ms), representing the time taken to process one image. Faster speeds, like the 80.4 ms achieved by the YOLOv8n model, indicate quicker inference times, making this variant suitable for applications where low latency is critical.
- **Speed A100 TensorRT (ms):** This column shows the inference speed when the models are deployed on an NVIDIA A100 GPU using TensorRT, a high-performance DL inference library. The latency is again measured in milliseconds. The YOLOv8n model exhibits an impressive speed of just 0.99 ms, demonstrating its capability for real-time applications on powerful GPU hardware.
- **Params (M):** The "Params" column refers to the number of parameters in each model, measured in millions (M). A higher number of parameters generally correlates with a model's capacity to learn more complex patterns but also increases computational demands. For instance, the YOLOv8x model, with 68.2 million parameters, has the highest capacity for detailed feature extraction, contributing to its higher mAP score.
- **FLOPs (B):** FLOPs, or Floating Point Operations, measure the computational complexity of a model, indicating how many billion (B) floating-point operations are required to make a prediction. Higher FLOPs typically mean that the model

is more computationally intensive, which could impact processing time, especially on less powerful hardware. The YOLOv8x model, with 257.8 billion FLOPs, is the most computationally demanding, reflecting its higher accuracy and larger number of parameters.

This detailed breakdown highlights the trade-offs between speed, accuracy, and computational requirements across the different YOLOv8 model variants.

TABLE 1. YOLOv8 Model Variants Performance Comparison (adapted from [1])

Model	Size (pixels)	mAP 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

3.2. Proposed Processes architecture

To address the challenge of accurately detecting both Persons and PPE in various environments, three distinct processes were developed based on the YOLOv8 model.

Unified Detection Process (UDP) employs a unified detection strategy, where a single YOLOv8 model is trained to simultaneously detect both persons and PPE items. The YOLOv8M variant was chosen for its balance between speed and accuracy, making it well-suited for handling the complexity of detecting multiple objects while maintaining efficient performance.

Parallel Detection Process (PDP) adopts a parallel detection strategy, where two distinct YOLOv8 models are used. First, the default YOLOv8L model is employed to locate individuals within the image. Then, a YOLOv8M model, which provides a balance between speed and accuracy, is used to detect PPE items such as boots, gloves, hats, and more.

Sequential Detection Process (SDP) follows a two-stage detection process similar to PDP but incorporates an additional image cropping step to improve precision. In this process, YOLOv8L is used for initial Person detection. Once the Persons are identified, the process crops out individual Persons from the image based on their bounding boxes, creating smaller, person-focused images. These cropped images are then processed by a YOLOv8M model trained to detect PPE items.

Together, these processes offer a variety of strategies to ensure reliable person and PPE detection, each with its advantages based on the complexity of the detection task. The following subsection will give more insights regarding each Process' architecture.

3.2.1. Unified Detection Process (UDP)

The first process utilises a unified approach, employing a single YOLOv8 model (specifically, the YOLOv8M variant) trained to detect both persons and various PPE items. This model is capable of identifying multiple objects within an image simultaneously, including

persons, boots, ear muffs, gloves, hats, protective glasses, and vests. By training a single model to recognise all these objects concurrently, the process aims to make the detection process more efficient, offering a comprehensive solution that simplifies the overall workflow.

Figure 5 illustrates the workflow of UDP. The process begins with the input image, which is fed into the YOLOv8M model. The model then processes the image, performing object detection to identify and localise all relevant classes, including both persons and various PPE items. The output is an annotated image where each detected object is marked with bounding boxes and labels, indicating the type of object detected.

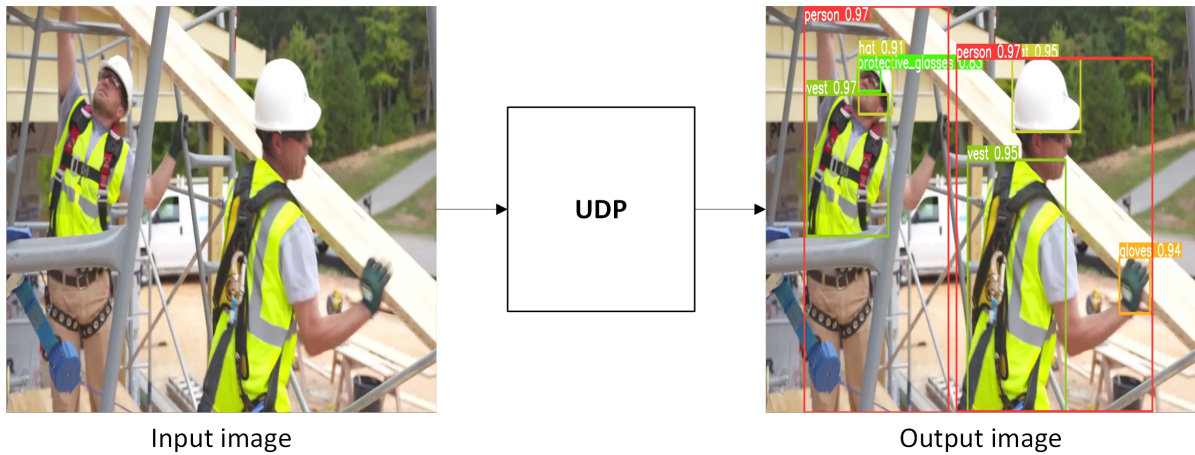


FIGURE 5. Workflow of UDP from input image to output with detected objects

3.2.2. Parallel Detection Process (PDP)

The second process adopts a parallel detection strategy using two distinct YOLOv8 models that process the entire image:

- **Person Detection:** The first stage employs YOLOv8 with default weights (YOLOv8L) specifically parameterised for only detecting persons. This model scans the entire image to locate and mark bounding boxes around individuals.
- **PPE Detection:** In the second stage, a trained YOLOv8 model (PPE_all) is used to detect PPE items such as boots, ear muffs, gloves, hats, protective glasses, and vests. This model also processes the entire image to identify PPE items.

This approach leverages the strengths of YOLOv8L in detecting persons and the PPE_all model's specialisation in identifying PPE items, ensuring a comprehensive detection of both persons and PPE across the entire image. Figure 6 illustrates the workflow of PDP from input to output.

Figure 6 demonstrates the parallel detection process of PDP. The process begins with the input image being fed into the YOLOv8L model for person detection. The detected persons are then marked with bounding boxes in the image. Subsequently, the same image is processed by the PPE_all model to detect various PPE items, which are also marked

with bounding boxes and labels. This two-step approach aims to improve the accuracy and reliability of detecting both persons and PPE items within the same image.

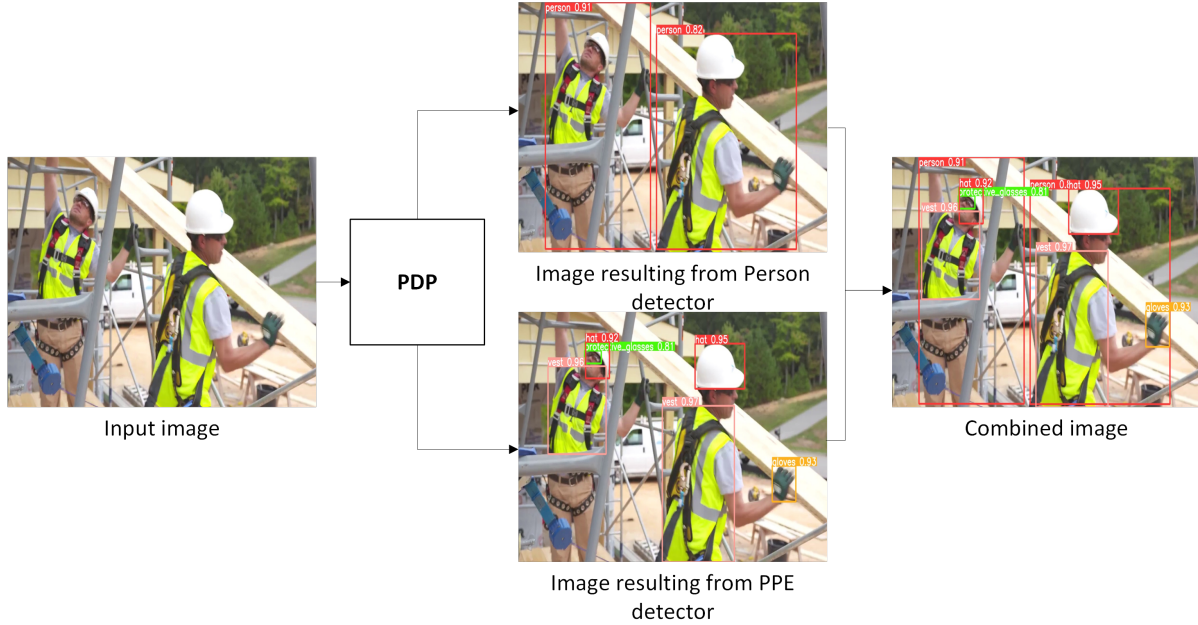


FIGURE 6. Workflow of PDP from input image to output with detected persons and PPE items

3.2.3. Sequential Detection Process (SDP)

The third process also utilises a two-stage detection process similar to PDP, but with an additional cropping step to enhance precision:

- **Person Detection:** The initial step uses YOLOv8 with default weights (YOLOv8L) to detect persons within the image, similar to PDP.
- **Image Cropping:** Once people are detected, the process crops these individuals out of the image based on their bounding boxes. This cropping isolates each person, creating smaller, focused images.
- **PPE Detection:** The cropped images are then fed into a YOLOv8 model (Cropped_PPE_all) trained to detect PPE items. By working on these smaller, person-specific images, the model can achieve higher accuracy in identifying PPEs such as boots, ear muffs, gloves, hats, protective glasses, and vests.

The cropping step in this process serves to reduce the complexity of the input for the PPE detection model, potentially improving detection performance by minimising the presence of irrelevant background elements. Figure 7 illustrates the workflow of SDP, from person detection to image cropping and final PPE detection.

These three process collectively offer a range of strategies for monitoring the use of PPEs using computer vision, each with its own advantages in terms of complexity, precision, and accuracy.



FIGURE 7. Overview of SDP: A two-stage detection process with image cropping for enhanced precision

3.3. Data Preparation and Assessment Metrics

In this section, it will be explained how the process of Data Preparation was made as well as it will be explained the metrics that were used to assess the models that were trained and the processes developed. For the data preparation phase, a selection of videos encountered on the internet was made to retrieve images for training and testing the processes. Four videos were chosen for this task, with two videos allocated only for training and validation purposes and the remaining two only for testing.

3.3.1. Training and Validation Data

To gather training data, a simple script using OpenCV2 [46] was developed. This script facilitated the conversion of video frames into individual images. Frames were sampled at intervals of 10 frames, resulting in approximately 3 images per second, ensuring a diverse representation of frames.

During the execution of the script, users interacted with it by pressing the 's' key. Upon pressing 's', the script initiated transforming video frames into images, capturing the frames at the specified intervals. This functionality provided users with control over the data collection process, allowing them to start and stop image capture as needed. Subsequent presses of the 's' key stopped the image capture process, providing users with flexibility in selecting frames for inclusion in the dataset. The idea behind this script was to ensure that only pictures with persons were saved and later used for model's training. The total number of images collected was 746, drawn from various segments of the videos. This sampling strategy ensured a comprehensive dataset encompassing a range of scenarios and conditions.

The distribution of the objects across the dataset, corresponding to different PPE-related classes, is as follows:

- **Person:** 1569 instances
- **Boots:** 874 instances
- **Ear Muffs:** 17 instances
- **Gloves:** 554 instances
- **Hat:** 1390 instances
- **Protective Glasses:** 602 instances
- **Vest:** 924 instances

This distribution highlights the varied presence of different PPE items in the dataset. While certain classes, such as ear muffs, have relatively fewer occurrences, other items, such as hats and vest, are more prevalent. This variation in class distribution provides valuable insights for model training, as it underscores the need for an incremental approach, that is, firstly the training of the first models should focus on classes with more objects such as hats and vests, following by gloves and boots, and finally protective glasses and ear muffs. These last two classes represent a challenge for the model’s training since the object corresponding to these classes may vary in form and it depends on how the person is positioned within the image.

Additionally, the Labelling tool [47] was used to annotate the images by drawing bounding boxes around PPE items used by persons. A bounding box is a rectangular border that encompasses an object in an image, providing spatial coordinates that indicate the object’s location. PPE items not used by persons were not considered in this task.

Following data collection, the dataset was partitioned into training and validation sets using a 70/30 split. This allocation was chosen to strike a balance between providing an wide amount of data for model training and retaining a separate set for validation. The 70% portion allocated to training facilitated the development of robust models, while the remaining 30% reserved for validation allowed for the evaluation of model performance on unseen data. This split was considered appropriate for preventing over-fitting and ensuring the generalisation of the trained models to unseen data.

3.3.2. Testing Data

For the testing phase, 50 random frames were selected from each of the two videos kept exclusively for testing purposes. This approach was taken to prevent bias towards frames where the models might perform better. Additionally, these frames were manually reviewed to ensure they were representative of the types of images used during training.

To assess the process’ performance across different scenarios, one video was selected that theoretically presented simpler conditions, making it easier for the process to detect correctly. Additionally, another video was chosen that presented more challenging conditions, posing greater difficulties for the process in accurately detecting Persons and PPE

items. The designations given to these datasets were **Test Data 1** and **Test Data 2**, respectively.

Similar to the training data, the Labelling tool was used to annotate the testing images by drawing bounding boxes around the PPE items.

3.3.3. Assessment metrics

For the evaluation of the trained models, it was used metrics such as Precision, Recall, F1-Score, Mean Average Precision at a confidence threshold of 0.5 (mAP50) and also Mean Average Precision across a range of confidence thresholds from 0.5 to 0.95 (mAP50-95). Before delving into the metrics, it's crucial to define the fundamental concepts of True Positive (TP), False Positive (FP), and False Negative (FN).

A TP occurs when the model correctly identifies an object that is actually present in the image. For example, if the model detects a hat on a worker and there truly is a hat in the image, this counts as a TP.

A FP happens when the model incorrectly identifies an object that is not present in the image. For instance, if the model detects a hat where there is none, this prediction is considered an FP.

A FN happens when the model fails to detect an object that is actually present in the image. If the model does not detect a hat on a worker even though one is visible, this would be considered an FN.

Precision measures the accuracy of the positive predictions. It is calculated as the ratio of TP predictions to the total number of positive predictions, which includes both TP and FP. The formula for Precision is shown in Equation 1.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

Recall measures the ability of the model to find all relevant instances in the dataset. It is calculated as the ratio of true positive predictions to the total number of actual positive instances, which includes both TP and FN. The formula for Recall is shown in Equation 2.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

The F1-Score is a harmonic mean of Precision and Recall, providing a balanced metric when there is an uneven class distribution. It is useful when both false positives and false negatives need to be minimised. The F1-Score is calculated using the formula shown in Equation 3.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Regarding mAP50, it measures the precision at a single threshold (50% overlap), while mAP50-95 averages precision over multiple thresholds (from 50% to 95% overlap).

IoU measures the overlap between predicted and ground truth bounding boxes and is calculated using the following equation:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (4)$$

where:

- **Area of Overlap:** The area where the predicted bounding box and the ground truth bounding box intersect.
- **Area of Union:** The total area covered by both the predicted and ground truth bounding boxes combined.

Figure 8 illustrates how IoU is calculated.

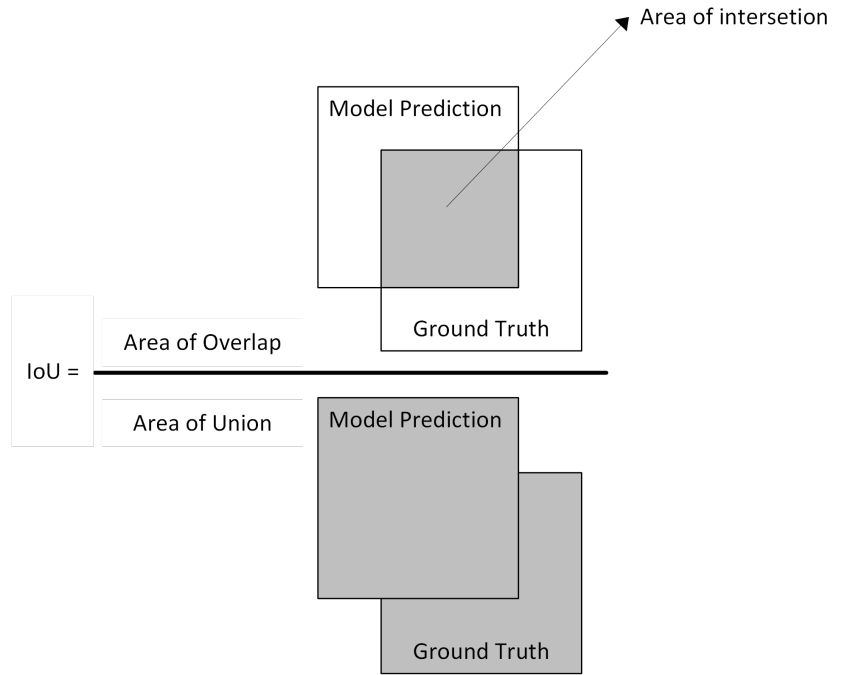


FIGURE 8. IoU calculation

During the training of the YOLOv8 model, several key loss metrics are also monitored to assess the model's learning progress and its ability to generalise to unseen data. These metrics include:

- **Box Loss (train/box_loss and val/box_loss):** This loss measures the accuracy of the predicted bounding boxes compared to the actual object locations in both the training and validation datasets. Lower values indicate that the model is becoming better at accurately locating objects within the images.
- **Classification Loss (train/cls_loss and val/cls_loss):** This metric evaluates how well the model classifies each detected object into the correct category. A reduction in classification loss signifies an improvement in the model's ability to correctly identify objects.

- **Distribution Focal Loss (train/dfl_loss and val/dfl_loss):** This loss focuses on the precision of the predicted bounding box edges. Lower values of distribution focal loss indicate that the model is enhancing its accuracy in predicting the exact boundaries of objects.

This comprehensive data preparation process ensured that the processes were trained and tested on diverse and accurately annotated datasets, allowing for robust evaluation of their performance in detecting and identifying PPE items.

CHAPTER 4

Training of Models

4.1. All Classes Detection (Person & PPE)

The initial phase of model training focused on training the YOLOv8 model to detect multiple classes, including persons and various PPE items such as boots, ear muffs, gloves, hats, protective glasses, and vests. The chosen model variant for this task was YOLOv8M because of its balance between speed and accuracy, making it well-suited for handling the complexity of detecting multiple objects while maintaining efficient performance. Training was conducted for 100 epochs to allow the model to learn patterns and features associated with each class. The model was trained to concurrently recognise and locate several objects inside an image by including many classes into the detection job. This approach offers a comprehensive solution for monitoring the appropriate use of PPEs.

The model demonstrated good generalisation during training, as evidenced by the consistent decrease in all loss values, including the box loss, classification loss, and distribution focal loss. As shown in Figure 9, these losses steadily approached zero throughout the training phase, indicating that the model was effectively learning to accurately detect and classify objects while minimising errors in bounding box predictions. This trend suggests that the model is not only fitting the training data well but is also likely to perform effectively on unseen data.

However, the precision values showed some instability, hovering between 80% and 90%, which might suggest occasional difficulty in consistently identifying specific classes or objects within the images. Despite this, the metrics such as recall, mAP50, and mAP50-95 showed continuous improvement throughout the model's training. Notably, the model trained during the 92nd epoch achieved the best metrics, indicating an optimal balance between under-fitting and over-fitting at this stage.

TABLE 2. Results of YOLOv8 Model Training for Detection of All Classes (Results expressed in percentage)

Model	Precision	Recall	F1-Score	mAP50	mAP50-95
All.classes	92.3	70.7	80.1	74.7	57.7

Table 2 summarises the performance metrics achieved by the YOLOv8M model trained for the detection of all classes. The Precision score of 92.3% indicates the accuracy of positive detections made by the model, while the Recall score of 70.7% measures the model's ability to correctly identify all relevant instances of the classes. The F1-Score of 80.1% reflects a balanced measure of Precision and Recall, providing an overall performance metric that considers both false positives and false negatives. The mAP50 scores of 74.7% and

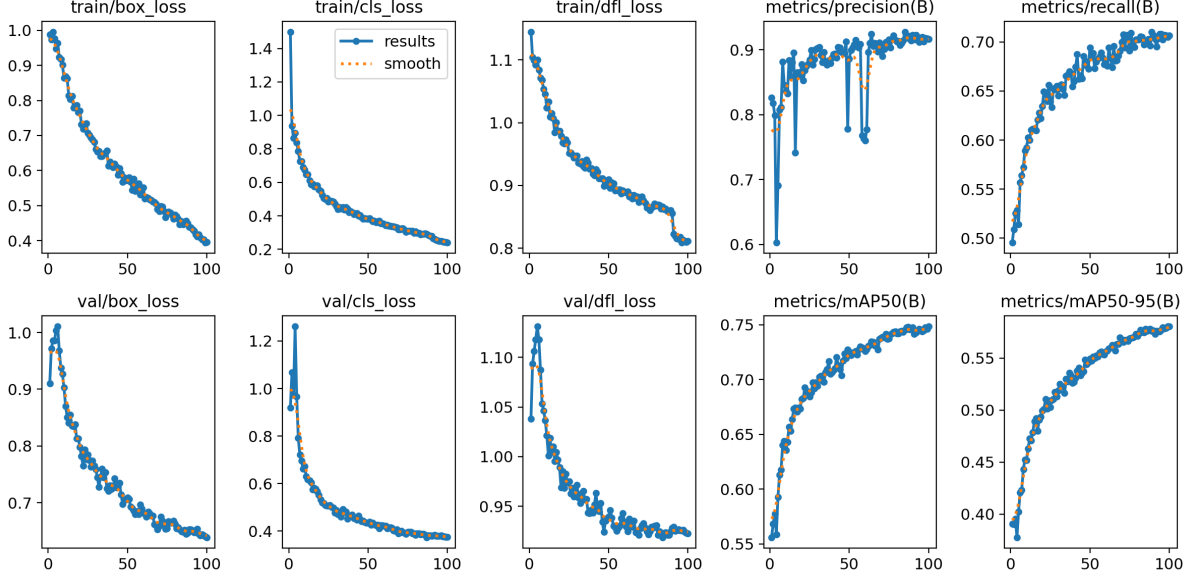


FIGURE 9. Loss and metrics curves during the training of All_classes model

57.7% for mAP50-95 reflect the model’s performance across different confidence thresholds, with higher scores indicating better performance. Overall, the YOLOv8M model demonstrates satisfactory performance in detecting all classes, achieving a balance between precision and recall, while maintaining a strong F1-Score.

4.2. Person detection

The initial phase of model training focused on training YOLOv8 models specifically for detecting persons. This step was crucial as it laid the foundation for subsequent training tasks involving the detection of Personal Protective Equipment (PPE) items.

To ensure accurate detection of persons, it was imperative to prepare the training data carefully. The training dataset consisted of images sampled from the collected dataset, ensuring a diverse representation of persons in various scenarios and conditions. These images were organised to exclude any bounding boxes corresponding to PPE items, thereby focusing solely on the detection of persons.

The training process involved training YOLOv8 models for 100 epochs, using two different variants: YOLOv8S and YOLOv8M. Each variant offered a different balance between model size, speed, and accuracy, allowing for experimentation to identify the optimal configuration for person detection.

The YOLOv8S demonstrated excellent generalisation during training, as evidenced by the consistent decrease in all loss values, including the box loss, classification loss, and distribution focal loss. As shown in Figure 10, these losses steadily approached zero throughout the training phase, indicating that the model was effectively learning to accurately detect persons while minimising errors in bounding box predictions. This consistent decrease in loss values suggests that the model is well-tuned and is likely to perform effectively on unseen data.

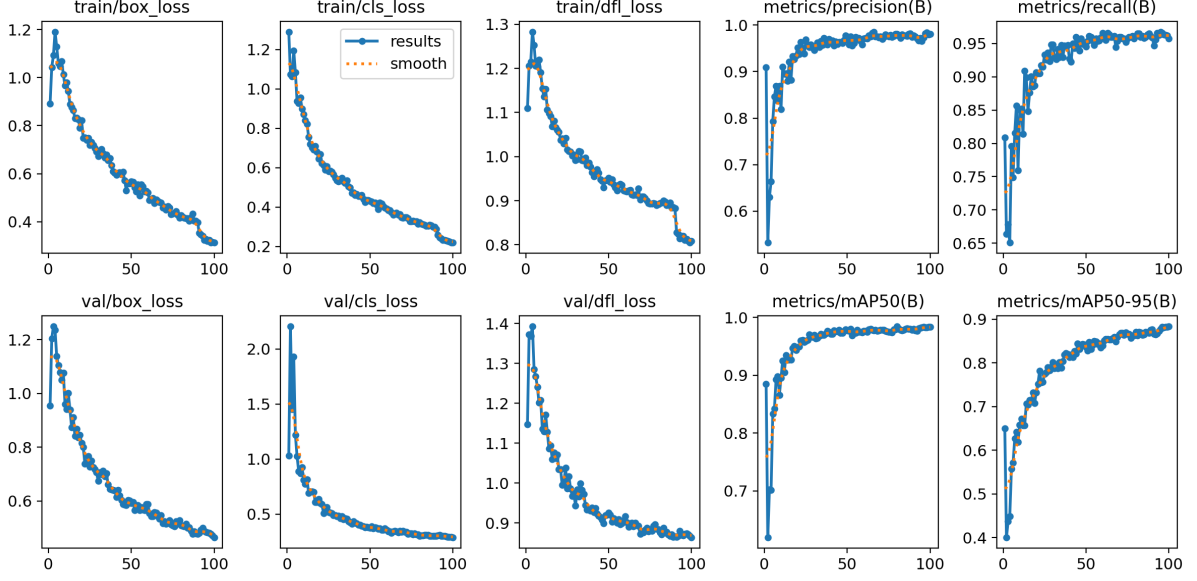


FIGURE 10. Loss and metrics curves during the training of Person model using YOLOv8S

Unlike the All_classes model, the precision metric for this person detection model remained consistently high throughout the training process. This stability in precision indicates that the model was consistently accurate in its positive classifications, successfully identifying persons in the images with a high degree of confidence.

Metrics such as recall, mAP50, and mAP50-95 also showed continuous improvement during the model's training, with the best overall performance observed in the 88th epoch. At this point, the model achieved an optimal balance between precision and recall, making it highly effective for the task of person detection.

The training of the YOLOv8M model revealed strong generalisation capabilities, as indicated by a steady reduction in all loss metrics, including box loss, classification loss, and distribution focal loss. As illustrated in Figure 11, these loss values consistently trended towards zero throughout the training process, highlighting the model's ability to refine its classifications and minimise errors in object detection. This downward trend in losses suggests that the model was effectively learning to identify and locate objects with increasing accuracy, which bodes well for its performance on new, unseen data.

In contrast to models that exhibited some variability, the precision of the YOLOv8M model remained consistently high from the beginning to the end of the training. This stability indicates that the model was reliable in making accurate positive classifications, consistently identifying objects with a high level of confidence.

Moreover, performance metrics such as recall, mAP50, and mAP50-95 improved steadily during the training period, with the model reaching its peak performance at the 68th epoch. At this point, the YOLOv8M model demonstrated an excellent balance between precision and recall, making it particularly well-suited for the tasks it was trained to perform.

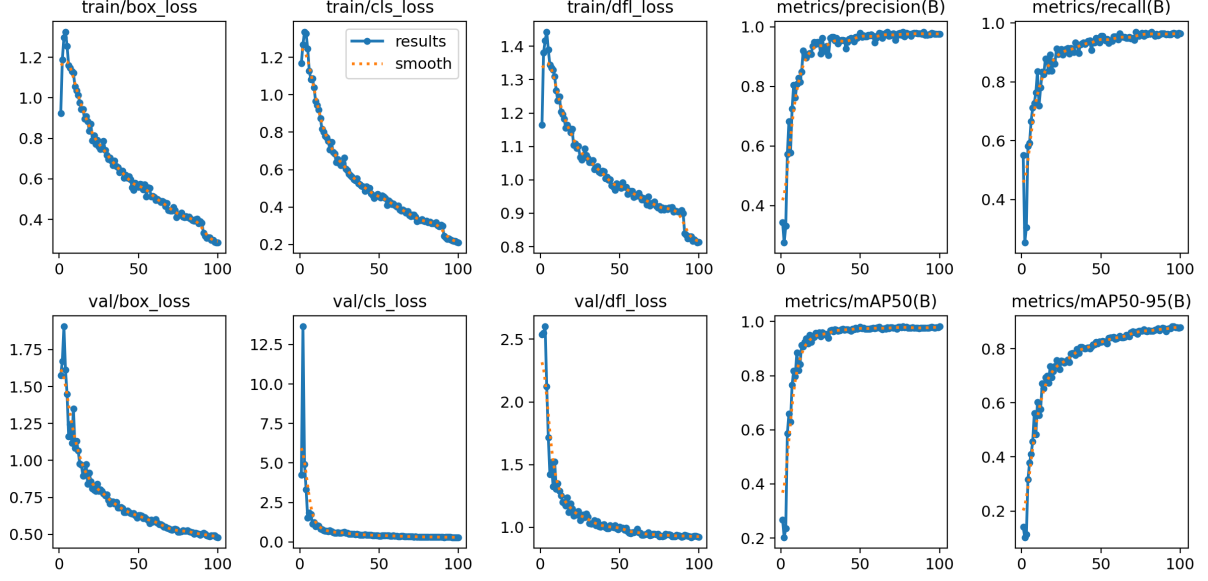


FIGURE 11. Loss and metrics curves during the training of Person model using YOLOv8M

TABLE 3. Results of YOLOv8 Models Trained for Person Detection (Results expressed in percentage)

Model	Precision	Recall	F1-Score	mAP50	mAP50-95
YOLOv8S	98.2	96.5	97.3	98.0	86.9
YOLOv8M	98.4	96.4	97.4	97.6	87.5

The results demonstrate in Table 3 that both YOLOv8S and YOLOv8M models achieved high Precision and Recall scores, indicating their effectiveness in accurately detecting persons in the images. The F1-Score, which balances Precision and Recall, is also very high for both models, with YOLOv8M slightly outperforming YOLOv8S. This indicates that YOLOv8M maintains a good trade-off between detecting all relevant instances (Recall) and avoiding false positives (Precision). Additionally, the mAP scores suggest that both models perform well across a range of confidence thresholds, with YOLOv8M slightly outperforming YOLOv8S in terms of mAP50.

Although the results were good, it was also evaluated the performance of YOLOv8 with the default weights in order to detect the class person. For that, it was used two variants of the model - YOLOv8M and YOLOv8L -, and it was also used the Ultralytics built-in task (validation).

TABLE 4. Results of YOLOv8 Models Validated for Person Detection (Results expressed in percentage)

Model	Precision	Recall	F1-Score	mAP50	mAP50-95
YOLOv8M	96.2	88.3	92.1	95.1	82.9
YOLOv8L	96.3	85.5	90.6	95.8	84.6

Although the results of the trained models were better, it was decided to keep the default weights upon visual verification. This decision was made because the performance of the models with default weights, shown in Table 4 was considered sufficient for the task at hand, and using the default weights avoided the need for additional training time and computational resources. On the other hand, it was tested the trained model in a Test dataset specific for the Person detection class and it was clear that the model had some difficulties in generalising for images different from the ones it was trained with.



(A) Detection result of the default YOLOv8 model, which successfully detects the person (highlighted in a red bounding box).



(B) Detection result of the trained YOLOv8 model, which fails to identify the person.

FIGURE 12. Comparison of person detection performance between the default YOLOv8 and the trained YOLOv8 model.

To illustrate the performance differences between the trained YOLOv8 model and the default YOLOv8 version, we present a visual comparison in Figure 12. The figure highlights two scenarios: in the upper part, the default YOLOv8 version successfully detects the person, as indicated by the red bounding box surrounding the individual, in the lower part, the trained YOLOv8 model fails to identify the person present in the image, whereas. This visual comparison underscores the performance disparity between the models, demonstrating the default YOLOv8's superior capability in detecting individuals under the same conditions.

4.3. PPEs detection

In this part, three different approaches were employed to train object detectors for the detection of PPE. The goal was to enhance the model's ability to accurately identify PPE items such as boots, ear muffs, gloves, hats, protective glasses, and vests. The following approaches were utilised:

1. Whole Image Training (PPE_xxx) : In the first approach, the entire image was used for training the object detectors. This method allows the model to learn from the context of the whole scene, capturing both the person and the surrounding environment. Figure 13 shows an example of an image used in the training of such models.



FIGURE 13. Example of image used in the training of the models

2. Cropped Images Training (Cropped_xxx) : The second approach involved cropping the images to only include the regions corresponding to the ground truth bounding boxes of the persons. By focusing on these regions, the model was trained specifically on the areas where the PPE items are most likely to be found, potentially improving the precision of PPE detection. Figure 14 shows an example of an image used in the training of such models.

3. Masked Background Training (Masked_xxx) : In the third approach, all areas outside the ground truth bounding boxes of the people were masked out (painted black) in the original images. This technique emphasises the regions containing the persons while ignoring the irrelevant background information. The objective was to improve the detection accuracy of PPE items on the subjects and minimise background noise by training the model on these masked photos. Figure 15 shows an example of an image used in the training of such models.



FIGURE 14. Example of a cropped image used in the training of the models



FIGURE 15. Example of a masked image used in the training of the models

Each of these approaches was designed to explore different aspects of the training data, with the goal of improving the model's ability to accurately detect and identify PPE items in various contexts. Additionally, the training of the models was divided into phases for each approach. Initially, the focus was on detecting the easier PPE items

such as hats and vests (XXX_hat_vest). Subsequently, the models were trained to detect boots and gloves (XXX_hat_vest_boots_gloves), and finally, the training included the more challenging items, protective glasses and ear muffs (XXX_all). This phased approach aimed to incrementally improve the model’s detection capabilities by starting with simpler tasks and progressively tackling more complex ones. For the training of such models, it was also considered the YOLOv8M version and the models were all trained for 100 epochs.

Starting off with the first approach, **Whole Image Training**, the models were trained progressively by introducing additional PPE classes. Initially, the focus was on training the model to detect simpler PPE classes such as hats and vests. Subsequently, the complexity of the models was increased by incorporating more classes, including boots and gloves, and finally including ear muffs and protective glasses as PPE classes. For data preparation, the training data excluded the "Person" class along with its corresponding bounding boxes.

The YOLOv8M model trained for detecting all PPE classes exhibited good generalisation during training, as evidenced by the consistent decrease in loss values, including the box loss, classification loss, and distribution focal loss. As illustrated in Figure 16, these losses progressively approached zero, indicating the model’s effectiveness in learning to detect various PPE items with minimal errors in bounding box predictions.

However, the precision metric for this PPE detection model showed a bit of inconsistency throughout the training process. This fluctuation in precision suggests that while the model was generally accurate in its classifications, there were occasional variations in its confidence when identifying PPE items in the images.

Other metrics, such as recall, mAP50, and mAP50-95, demonstrated continuous improvement during the training, with the best overall performance observed at the 98th epoch. At this point, the model reached its optimal balance between precision and recall, indicating its heightened effectiveness in detecting and correctly classifying all PPE items.

TABLE 5. Results of YOLOv8 Models Training for PPE Detection. (Results expressed in percentage)

Model	Precision	Recall	F1-Score	mAP50	mAP50-95
PPE_hat_vest	98.1	94.0	96.0	97.0	83.6
PPE_hat_vest_boots_gloves	90.8	78.9	84.4	84.9	64.7
PPE_all	91.7	65.3	76.3	69.7	52.3
Cropped_PPE_hat_vest	97.6	93.3	95.4	96.7	84.1
Cropped_PPE_hat_vest_boots_gloves	91.6	83.3	87.3	88.0	68.2
Cropped_PPE_all	93.2	83.8	88.3	88.8	60.9
Masked_PPE_hat_vest	97.7	93.2	95.4	96.6	83.6
Masked_PPE_hat_vest_boots_gloves	90.2	80.7	85.2	84.6	64.0
Masked_PPE_all	91.6	65.0	76.0	69.8	52.3

The results in Table 5 demonstrate that as the model’s complexity increased, it maintained robust performance across different metrics. Notably, the PPE_hat_vest model achieved high Precision (98.1%) and Recall (94.0%), resulting in an F1-Score of 96.0%.

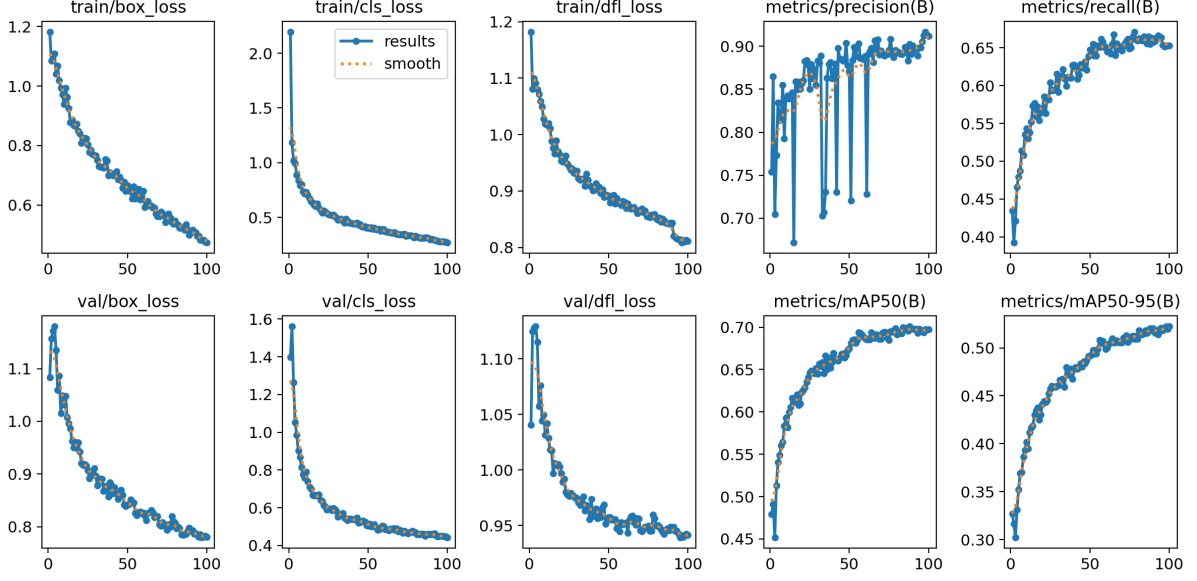


FIGURE 16. Loss and metrics curves during the training of PPE_all model using YOLOv8M

This indicates the model’s excellent ability to accurately detect and classify hats and vests. As additional classes were introduced in the PPE_hat_vest_boots_gloves and PPE_all models, a decrease in performance metrics was observed, especially in Recall, which suggests that detecting more PPE items poses additional challenges and affects overall model performance.

The performance decline in recall for the PPE_all model suggests challenges in detecting smaller or less distinct PPE items like protective glasses and ear muffs. This could be attributed to their variability in appearance and occlusion by other objects or body parts. However, the models consistently maintained high precision, indicating low false positive rates across all experiments.

Moving to the second approach, **Cropped Ground Truth Training**, significant adjustments were made to the dataset preparation process. In this method, each image suffered a transformation where the person of interest was cropped based on their ground truth bounding box. This approach was crucial to ensure that the training focused specifically on the regions containing persons and their associated PPEs.

To adapt the dataset for training under this approach, every image was processed to isolate the person of interest along with their PPE items. This involved extracting the relevant portion of the image defined by the ground truth bounding box of the person. Subsequently, all coordinates related to the PPE annotations were adjusted to align with the new cropped image dimensions.

The YOLOv8M model trained on the Cropped_PPE_all dataset showed commendable progress throughout the training phase, as illustrated by the gradual decrease in loss values, including box loss, classification loss, and distribution focal loss. However, it is noteworthy that around the 50th epoch, the validation distribution focal loss (val/df_l_loss)

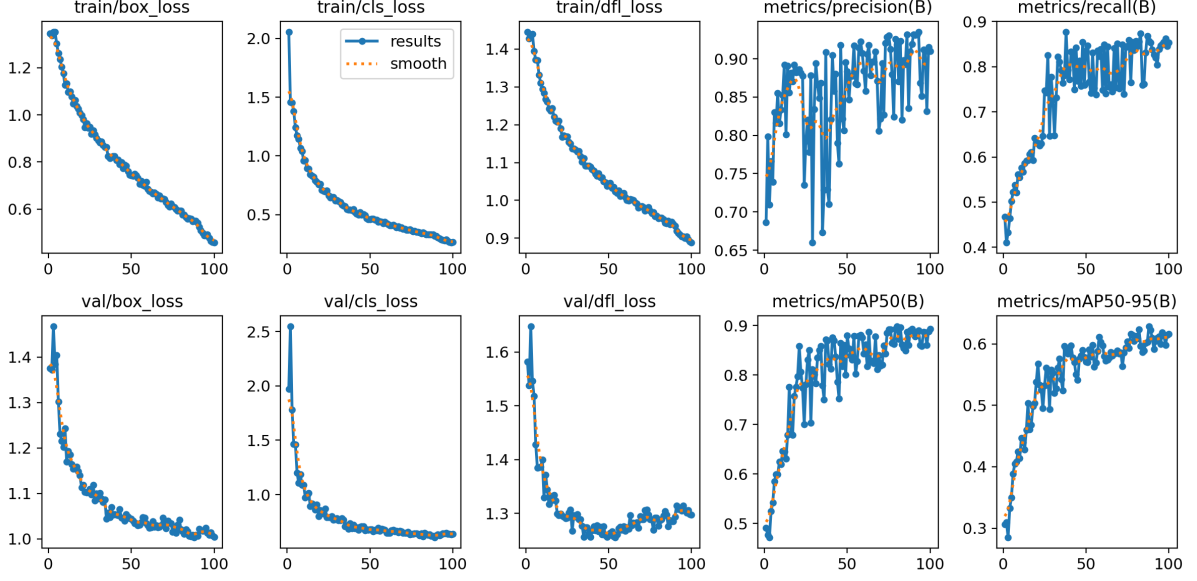


FIGURE 17. Loss and metrics curves during the training of Cropped_PPE_all model using YOLOv8M

started to increase, indicating some challenges in maintaining the model’s performance as training progressed. Despite this, as depicted in Figure 17, the overall trend in loss values remained downward, suggesting that the model was effectively learning to detect and classify all PPE items while minimising errors in bounding box predictions.

However, the precision and recall metrics for this model displayed some inconsistencies during the training process. These fluctuations suggest that while the model was generally adept at detecting PPE items, there were occasional variances in its ability to consistently identify and classify all instances within the images accurately.

The mAP50 and mAP50-95 metrics also exhibited some inconsistency, though these variations were not as pronounced as those observed in precision and recall. Despite these minor fluctuations, the model achieved its best overall performance at the 92nd epoch, where it reached an optimal balance between precision and recall, resulting in highly effective detection of all specified PPE items.

The results in Table 5 demonstrate the effectiveness of training YOLOv8 models on cropped images. For instance, the Cropped_PPE_hat_vest model showed high Precision (97.6%) and Recall (93.3%) with an F1-Score of 95.4, which is a notable improvement over the original approach. This method, which cropped images to focus on the person and their PPE items, resulted in better overall performance metrics, highlighting the effectiveness of focusing on specific regions of interest in training data. The Cropped_PPE_all model also demonstrated higher F1-Scores compared to the PPE_all model from the first approach, indicating enhanced detection capability with less background interference.

Moving to the third approach, **Masked Background Training**, the dataset preparation process involved a significant modification where all areas outside the ground truth bounding boxes of the persons were masked out (painted black) in the original images.

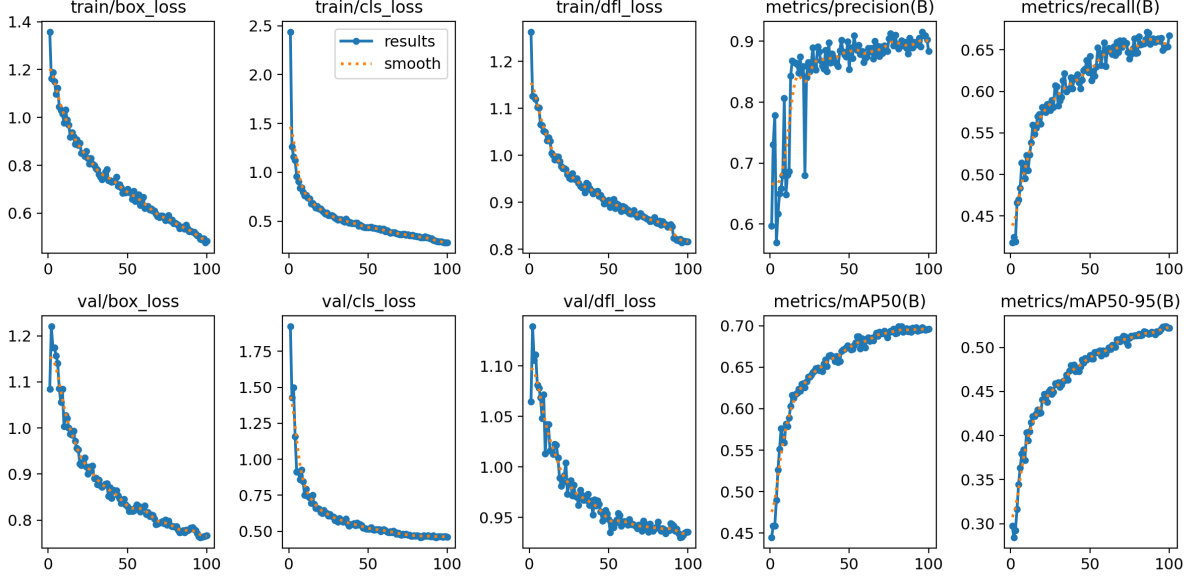


FIGURE 18. Loss and metrics curves during the training of Masked_PPE_all model using YOLOv8M

To prepare the dataset for training under this approach, each image experienced a transformation where only the portion inside the person’s ground truth bounding box was retained and the rest of the image was masked with black. This ensured that the model focused exclusively on the areas containing persons and their associated PPE items, without being distracted by the background context.

The YOLOv8M model trained on the Masked_PPE_all dataset demonstrated strong performance throughout the training phase, as reflected in the steady decrease in loss values, including box loss, classification loss, and distribution focal loss. This trend, shown in Figure 18, indicates that the model was effectively learning to detect and classify all PPE items with increasing accuracy.

In the early stages of training, the precision metric for this model exhibited some inconsistency, suggesting that the model initially struggled with consistently accurate positive classifications. However, as training progressed, precision began to increase steadily, improving epoch by epoch. This upward trend highlights the model’s growing capability to accurately identify PPE items as it continued to refine its classifications.

While precision showed early fluctuations, other metrics, such as recall, mAP50, and mAP50-95, remained relatively stable and consistently improved over time. The model reached its peak performance at the 96th epoch, achieving the best balance between precision and recall, which resulted in highly effective detection of all specified PPE items.

For the Masked Background Training approach, as shown in Table 5, the model maintained high Precision (97.7%) but showed a decrease in Recall (93.2%) and F1-Score (95.4%) compared to the Cropped approach. While this method improved the focus on the PPE items by reducing background noise, the decrease in recall in the Masked_PPE_all model suggests that masking the background might have led to challenges in detecting all

instances of PPE items, particularly in more complex scenarios. This approach showed stable performance in mAP50 and mAP50-95 metrics but reflected a trade-off between precision and recall.

CHAPTER 5

Results and Discussion

To evaluate the performance of the processes, two different test datasets were used, each containing 50 images as mentioned previously. One of these datasets - **Test Data 1** - contains images that, in theory, are easier for the processes to detect correctly and the other one - **Test Data 2** - has more challenging images to classify. The metrics employed for evaluation were Precision, Recall, and F1-Score. These metrics provide a comprehensive understanding of the model's ability to accurately detect and classify PPE items. The inclusion of the F1-Score, which combines Precision and Recall into a single metric, is particularly useful in assessing the balance between these two measures, especially in scenarios where either false positives or false negatives are more prevalent.

For the evaluation of **UDP**, which involves the single-process detection, precision, recall and F1-Score were calculated by directly comparing the predicted bounding boxes to the ground truth. This straightforward approach allowed for an immediate assessment of the model's accuracy in identifying PPE items. However, for the other two processes, **PDP** and **SDP**, a more demanding evaluation criteria was applied. In these processes, a TP was only considered valid if the PPE item was correctly detected within a person who was also correctly detected. This means that if a person was not detected correctly but the PPE was, it was counted as a FN. This evaluation method ensures that the models not only detect PPE items accurately but also associate them correctly with the detected persons, enhancing the reliability of the detection process.

5.1. Unified Detection Process (UDP)

For the unified detection process, the model that was chosen is the **All_classes** model, which was discussed in the previous chapter. This model was selected due to its overall strong performance across various metrics, making it a reliable choice for detecting all PPE classes simultaneously.

TABLE 6. Performance of Unified Detection Process for each test dataset
(Results expressed in percentage)

Test Dataset	Precision	Recall	F1-Score
Test Data 1	73.0	45.7	56.2
Test Data 2	53.7	29.3	37.9

The results in Table 6 illustrate the performance of the All_classes model when applied to two different test datasets. For **Test Data 1**, the model achieved a precision of 73.0% and a recall of 45.7%, resulting in an F1-Score of 56.2%. This F1-Score reflects a moderate balance between precision and recall, indicating that while the model is reasonably

		Confusion Matrix							
Actual	person	81	0	0	0	0	0	0	10
	boots	0	3	0	0	0	0	0	13
	ear_muffs	0	0	0	0	0	0	0	0
	gloves	0	1	0	3	0	0	0	3
	hat	0	0	0	0	71	0	0	9
	protective_glasses	0	0	0	0	0	0	0	2
	vest	0	0	0	0	1	0	66	10
	background	6	10	0	1	1	2	2	0
		person	boots	ear_muffs	gloves	hat	protective_glasses	vest	background
		Predicted							

FIGURE 19. Confusion Matrix for UDP on Test Data 1

accurate in identifying PPE items and persons, it still misses a significant number of true positives, as seen in the recall. The F1-Score thus captures this trade-off, suggesting that while precision is higher, the recall drops enough to affect the overall effectiveness of the model.

For **Test Data 2**, the precision was 53.7%, indicating a lower accuracy in correctly identifying PPE items and persons compared to Test Data 1. The recall decreased further to 29.3%, which brought the F1-Score down to 37.9%. This low F1-Score highlights the process' struggle in balancing precision and recall in more challenging images, where variations in conditions like lighting, occlusion, and environmental complexity significantly impacted its performance. The sharp drop in F1-Score compared to Test Data 1 reflects that the model's ability to detect both persons and PPE is less reliable under difficult conditions, especially in terms of recall.

The confusion matrix in Figure 19 illustrates the effectiveness of the UDP on Test Data 1. Classes such as Person, hat and vest were the most correctly identified indicating the capability of the Process to detect these classes correctly. On the other hand, the class boots is the one that the Process failed to identify the most.

		Confusion Matrix							
Actual	person	34	0	0	0	0	0	1	29
	boots	0	5	0	0	0	0	0	26
	ear_muffs	0	0	0	0	0	0	0	0
	gloves	0	0	0	0	0	0	0	9
	hat	0	0	0	0	33	0	0	27
	protective_glasses	0	0	0	0	0	5	0	4
	vest	0	0	0	0	0	0	15	12
	background	4	4	0	2	0	0	1	0
		person	boots	ear_muffs	gloves	hat	protective_glasses	vest	background
		Predicted							

FIGURE 20. Confusion Matrix for UDP on Test Data 2

The confusion matrix in Figure 20 illustrates how the Process performed in Test Data 2. The Process achieved a relatively bad performance on this dataset, highlighting the need for further improvements on the model used.

Overall, the performance of the single-process detection using the All_classes model shows strengths in precision but underscores the need for improvements in recall. The F1-Score, particularly, highlights the model’s overall balance and emphasises areas for improvement, suggesting that future work could involve refining the model architecture, augmenting the training data, or incorporating additional pre-processing steps to address the challenges observed in recall, especially in more complex scenarios.

5.2. Parallel Detection Process (PDP)

In the second process configuration, the detection task is split into parallel steps: first detecting persons using YOLOv8L, and then detecting PPEs using the PPE.all model. This sequential approach aims to enhance the accuracy of PPE detection by focusing first on identifying persons and then specifically looking for PPE items on the detected individuals.

The results in Table 7 demonstrate a significant improvement in both precision and recall for PDP compared to UDP. For **Test Data 1**, PDP achieved a precision of 86.4% and a recall of 83.5%, resulting in an F1-Score of 84.9%. The higher F1-Score in PDP

TABLE 7. Performance of Parallel Detection Process for each test dataset
(Results expressed in percentage)

Test Dataset	Precision	Recall	F1-Score
Test Data 1	86.4	83.5	84.9
Test Data 2	87.1	64.0	73.8

indicates that splitting the detection process into person and PPE detection stages results in better balance and overall performance.

For **Test Data 2**, PDP maintained a high precision of 87.1%, but its recall dropped to 64.0%, resulting in an F1-Score of 73.8%. The decrease in recall in PDP shows that detecting PPE in more challenging images remains problematic. Nevertheless, the gap between precision and recall in PDP is smaller than in UDP, which demonstrates that PDP’s two-step process improves consistency in identifying PPE items in more difficult scenarios.

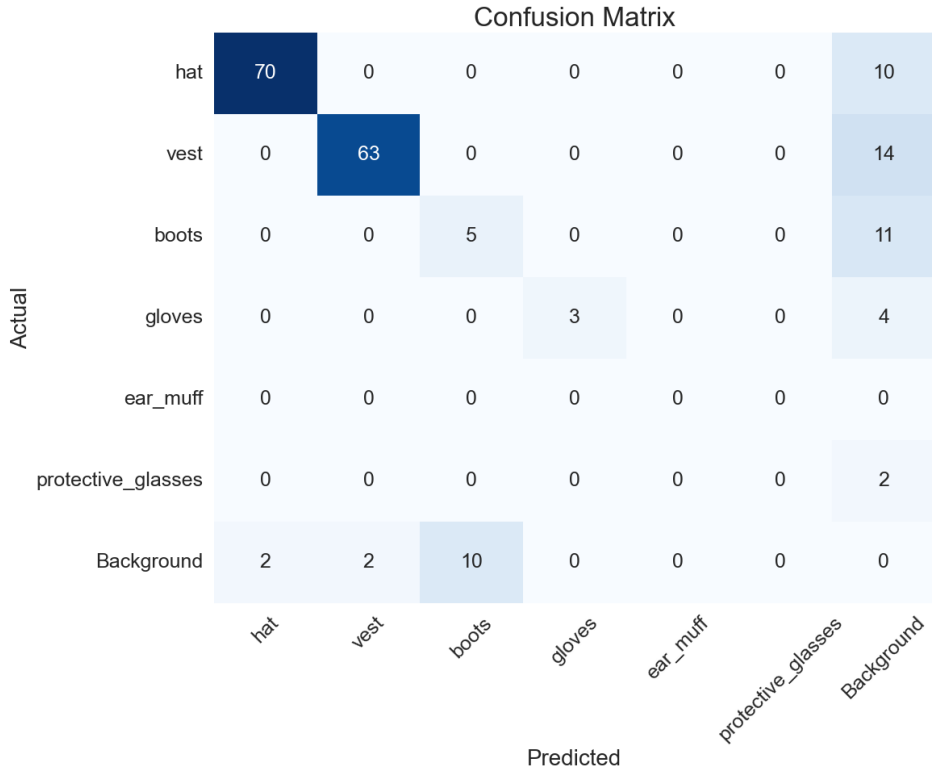


FIGURE 21. Confusion Matrix for PDP on Test Data 1

The confusion matrix in Figure 21 further highlights the effectiveness of PDP on Test Data 1, where the model correctly identified a high number of instances across most PPE categories, particularly hats and vests. The lower number of FP also suggests that the process is more reliable in maintaining precision. However, challenges remain with detecting certain items like boots and gloves, where some FP occurred, but these are still fewer compared to UDP.

		Confusion Matrix						
Actual	hat	36	0	0	0	0	0	24
	vest	0	19	0	0	0	0	8
	boots	0	0	9	0	0	0	22
	gloves	0	0	0	1	0	0	8
	ear_muff	0	0	0	0	0	0	0
	protective_glasses	0	0	0	0	0	5	4
	Background	3	0	2	1	0	0	0
		hat	vest	boots	gloves	ear_muff	protective_glasses	Background
		Predicted						

FIGURE 22. Confusion Matrix for PDP on Test Data 2

In **Test Data 2**, as shown in Figure 22, the precision remains high across most PPE categories, but recall drops, particularly for items like gloves and ear muffs, which were either misidentified or missed entirely. This reflects the challenges of more complex environmental conditions in Test Data 2. Despite these difficulties, PDP still significantly outperformed UDP, where recall was notably lower.

5.3. Sequential Detection Process (SDP)

In the third process configuration, the detection task is approached in sequential stages: first detecting persons using YOLOv8L, followed by detecting PPEs using the Cropped_PPE_all model. This approach focuses on cropping the detected persons from the image before applying the PPE detection, aiming to improve the detection of PPE items by focusing on person-specific regions.

TABLE 8. Performance of Sequential Detection Process for each test dataset (Results expressed in percentage)

Test Dataset	Precision	Recall	F1-Score
Test Data 1	82.9	76.2	79.4
Test Data 2	84.2	62.4	71.7

The results in Table 8 show that SDP achieves a balance between precision and recall, similar to PDP but with some differences. For **Test Data 1**, SDP achieved a precision of 82.9% and a recall of 76.2%, resulting in an F1-Score of 79.4%. This performance

is slightly lower than PDP in terms of precision and recall, but still significantly higher than UDP. The cropping step seems to help in reducing FP, as demonstrated by the high precision values, but it may also result in missed detections (lower recall) compared to PDP.

In **Test Data 2**, SDP maintained a high precision of 84.2%, comparable to PDP’s 87.1%, but its recall dropped to 62.4%, similar to PDP’s recall of 64.0%. The overall F1-Score of 71.7% is close to that of PDP (73.8%), indicating that both processes perform similarly in challenging environments. However, SDP’s slight drop in recall suggests that the cropping step may sometimes overlook parts of the person, leading to fewer PPE detections, especially in difficult scenarios.

		Confusion Matrix						
Actual	hat	162	0	0	0	0	0	17
	vest	0	49	0	0	0	0	31
	boots	0	0	6	0	0	0	12
	gloves	0	0	0	1	0	0	6
	ear_muff	0	0	0	0	0	0	0
	protective_glasses	0	0	0	0	0	0	2
	Background	18	1	10	3	0	9	0
		hat	vest	boots	gloves	ear_muff	protective_glasses	Background
		Predicted						

FIGURE 23. Confusion Matrix for SDP on Test Data 1

The confusion matrix in Figure 23 illustrates the performance of SDP on Test Data 1. The model shows high accuracy for detecting hats, vests, and boots, with no FP for these classes. However, similar to Processes 1 and 2, gloves, ear muffs, and protective glasses remain difficult to detect, with no instances detected for ear muffs or protective glasses in this dataset. This high precision, particularly in hats and vests, reflects the effectiveness of the cropping step in reducing false positives.

In **Test Data 2**, as seen in Figure 24, the performance of SDP remained consistent, with high precision for hats, vests, and boots. However, gloves and protective glasses had lower detection rates, and ear muffs were not detected at all. This mirrors the performance seen in PDP, where challenging PPE items like gloves and ear muffs proved difficult to detect. The background confusion, where non-PPE items are mistaken for PPE, is also

		Confusion Matrix						
Actual	hat	94	0	0	0	0	0	34
	vest	0	21	0	0	0	0	7
	boots	0	0	9	0	0	0	22
	gloves	0	0	0	2	0	0	7
	ear_muff	0	0	0	0	0	0	0
	protective_glasses	0	0	0	0	0	2	7
	Background	6	5	5	5	0	3	0
		hat	vest	boots	gloves	ear_muff	protective_glasses	Background
		Predicted						

FIGURE 24. Confusion Matrix for SDP on Test Data 2

reduced in SDP, indicating that the cropping step helps filter out irrelevant background elements.

5.4. Comparative Analysis with Previous Studies

Comparing the findings of this dissertation with prior research presents several challenges due to variations in datasets, object detector architectures, and evaluation criteria. Furthermore, not all studies employ the same set of PPE classes, which also complicates direct comparisons between their results. Nevertheless, a qualitative comparison can reveal some key trends and insights.

Previous studies in the construction, electrical, and industrial sectors have utilised various object detection architectures for PPE monitoring, including Faster R-CNN, SSD, and earlier versions of YOLO. For instance, [17] employed a Faster R-CNN architecture, achieving an accuracy and precision of around 80% in detecting incomplete PPE usage. Similarly, studies such as those by [23] and [20] utilised SSD-based models, achieving notable results in terms of precision and real-time processing capabilities. However, the mAP values reported in these studies, while significant, tend to be smaller than those achieved by more recent YOLO-based approaches.

In contrast, YOLO-based models, particularly from version 3 onwards, have demonstrated a strong balance between speed and accuracy. [24] achieved an impressive mAP of 93% using YOLOv3, and [31] reported a mAP of 86.6% with YOLOv5, highlighting the evolution of the YOLO architecture in improving detection capabilities. These studies

underscore the growing preference for YOLO architectures in the field of PPE detection, largely due to their efficiency in handling real-time scenarios.

More recent research has seen the adoption of YOLOv8, as evidenced by [32] and [33], who achieved precision values of 80.6% and 86%, respectively, across various PPE classes. These studies indicate that YOLOv8 continues the trend of enhancing detection accuracy and processing speed, making it a suitable choice for complex industrial environments.

In this dissertation, YOLOv8 was selected for its advanced capabilities in both detection accuracy and processing efficiency. The decision was based on a thorough literature review that identified YOLOv8 as a state-of-the-art architecture with the ability to process high-resolution images in real-time, crucial for effective PPE monitoring. The results obtained in this study are in line with the performance trends observed in other YOLO-based studies, affirming YOLOv8's superiority over earlier models like YOLOv3, YOLOv5, and even non-YOLO architectures such as Faster R-CNN and SSD.

Despite differences in datasets, PPE classes, and specific task requirements, the consistent use of mAP as a performance metric across studies facilitates a relative comparison. The findings from this dissertation demonstrate that YOLOv8 is competitive with, and in many cases superior to, non-YOLO architectures such as Faster R-CNN and SSD, particularly in real-time PPE detection contexts. This supports the broader trend of adopting the latest object detection models to enhance workplace safety across various industrial sectors.

CHAPTER 6

Conclusion

6.1. Final Remarks

The primary goal of this dissertation was to develop and evaluate processes that utilise DL models for monitoring the use of PPE, with a specific focus on the effectiveness of the YOLOv8 model. Throughout the research, the objectives were systematically pursued and successfully achieved. Additionally, the research questions were thoroughly explored, providing valuable insights into the application of DL models for improving workplace safety.

The first objective was to train and implement different object detectors, including the YOLOv8 model, to accurately detect and classify various types of PPE in real-world images. This was accomplished by successfully training the YOLOv8 model. The results demonstrated that YOLOv8 exhibited high accuracy in identifying and categorising PPE, confirming its effectiveness in real-world applications.

The second objective involved designing and developing integrated processes that leverage these DL models for PPE monitoring. This was achieved by building and testing a process that utilised the trained YOLOv8 model. The process was evaluated in practical scenarios and, despite challenges such as varying environmental conditions, it performed effectively in monitoring tasks. These findings highlight the potential for deploying such processes in real industrial environments.

The third objective was to explore the practical benefits of integrating person detection with PPE detection to build a more effective real-time PPE monitoring process across various industrial environments. The research demonstrated that combining these two detection tasks can significantly enhance workplace safety by ensuring both the accurate identification of individuals and the correct use of PPE. This integrated approach offers several advantages, including improved monitoring efficiency, reduced instances of missed PPE violations, and a more comprehensive adaptation to dynamic risks and challenges across different industries.

A comprehensive evaluation of the developed processes was also conducted, identifying strengths such as accuracy and real-time performance, along with areas for improvement, such as enhancing the robustness of detection in complex environments. This evaluation also led to the proposal of potential future advancements, including the integration of advanced DL and Computer Vision techniques to further enhance PPE monitoring capabilities.

Regarding the research questions: First, the effectiveness of the YOLOv8 model in detecting and classifying PPE in real-world scenarios was confirmed, as it achieved high

accuracy across some types of protective equipment. Even so, there are still some PPE classes, such as gloves, boots, protective glasses and ear muffs, that can receive some improvements. Second, the DL-based processes performed effectively in PPE monitoring, although challenges such as environmental variability and computational limitations were identified, indicating the need for further optimisation. Lastly, the study confirmed the benefits of integrating person detection with PPE detection, showing that combining these two tasks significantly enhances the effectiveness of monitoring processes. This integration improved the overall accuracy and reliability of detecting PPE usage, thus providing a more comprehensive approach to ensuring workplace safety by reducing safety violations and lowering the risk of accidents and injuries.

As a direct outcome of the research conducted in this dissertation, an article detailing the development and findings of the proposed PPE monitoring processes has been submitted to the journal *Multimedia Tools and Applications* for review and potential publication. The article can be found in the annexes for further reference.

In conclusion, this dissertation successfully achieved its objectives and provided a comprehensive exploration of DL-based processes for real-time PPE monitoring. The findings contribute to the growing body of knowledge in the field of Computer Vision and DL applied to workplace safety. The developed processes demonstrate the potential for practical deployment in various industrial environments, offering a promising pathway toward enhancing safety standards and mitigating risks in hazardous work conditions.

6.2. Future Work

Several promising directions could further advance this research. First, exploring the use of more recent versions of YOLO, such as YOLOv9 and YOLOv10, could offer improved accuracy and performance over the current model. Additionally, investigating other object detectors and comparing their performance in this context would provide valuable insights into which models are best suited for PPE monitoring tasks.

One of the directions that can be followed is the one that takes advantages of video data. One key area is the incorporation of tracking algorithms to continuously monitor individuals across multiple frames. This approach could enhance detection accuracy by mitigating the impact of momentary occlusions or detection failures in single frames. By integrating temporal information, tracking algorithms can help maintain consistent identification of individuals and their associated PPE over time. Additionally, employing a multi-frame analysis strategy could improve detection reliability. By evaluating a sequence of frames, the process can make more informed decisions and compensate for transient issues where PPE might not be detected in every single frame. This could lead to more robust monitoring systems capable of maintaining high accuracy and reliability even in challenging conditions.

Another potential area of future work is exploring a two-stage system: first employing an object detector to identify PPE, followed by a classifier to categorise those detected

objects. This approach could improve precision and offer greater flexibility in handling complex scenarios.

Furthermore, the integration of pose estimators offers another promising direction. Pose estimation could enhance the system’s ability to understand the spatial relationships between workers and their PPE, ensuring that not only is the PPE detected, but it is also worn correctly. Incorporating pose estimators could significantly improve the detection of incorrect or unsafe PPE usage, providing a more comprehensive solution to workplace safety.

Moreover, collecting a more diverse and extensive dataset will be crucial for better training and generalisation of the models. Expanding the dataset to include more variations in environments, lighting, and PPE types will help the models perform more robustly in a wider range of real-world conditions. This would further enhance the effectiveness and reliability of PPE monitoring systems in various industrial settings.

Bibliography

- [1] Ultralytics, “Home - ultralytics yolo docs,” 2024.
- [2] “Crisp-dm for data science teams: 5 actions to consider - data science process alliance.”
- [3] SafetyToolBox, “Personal protective equipment (ppe),” 2023.
- [4] “What is yolov8? the ultimate guide. [2024].”
- [5] Eurostat, “Accidents at work statistics,” 10 2023.
- [6] R. Sehsah, A. H. El-Gilany, and A. M. Ibrahim, “Personal protective equipment (ppe) use and its relation to accidents among construction workers,” *La Medicina del Lavoro*, vol. 111, p. 285, 7 2020.
- [7] H. Mohd Amir Shazwan and J. M. Ee, “Effectiveness of personal protective equipment (ppe) at construction site,” *Inti Journal*, vol. 1, no. 12, pp. 1–12, 2018.
- [8] R. Wirth and J. Hipp, “Crisp-dm: Towards a standard process model for data mining,” in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1, pp. 29–39, Manchester, 2000.
- [9] H. Snyder, “Literature review as a research methodology: An overview and guidelines,” *Journal of Business Research*, vol. 104, pp. 333–339, 11 2019.
- [10] J. M. G. de Gabriel, J. A. Fernández-Madrigal, M. del Carmen Rey-Merchán, and A. López-Arquillos, “A safety system based on bluetooth low energy (ble) to prevent the misuse of personal protection equipment (ppe) in construction,” *Safety Science*, vol. 158, p. 105995, 2 2023.
- [11] A. Kelm, L. Laußat, A. Meins-Becker, D. Platz, M. J. Khazaei, A. M. Costin, M. Helmus, and J. Teizer, “Mobile passive radio frequency identification (rfid) portal for automated and rapid control of personal protective equipment (ppe) on construction sites,” *Automation in Construction*, vol. 36, pp. 38–52, 12 2013.
- [12] A. Pisu, N. Elia, L. Pompianu, F. Barchi, A. Acquaviva, and S. Carta, “Enhancing workplace safety: A flexible approach for personal protective equipment monitoring,” *Expert Systems with Applications*, vol. 238, p. 122285, 3 2024.
- [13] X. Yang, Y. Yu, S. Shirowzhan, S. Sepasgozer, and H. Li, “Automated ppe-tool pair check system for construction safety using smart iot,” *Journal of Building Engineering*, vol. 32, p. 101721, 11 2020.
- [14] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 779–788, 6 2015.
- [16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016.
- [17] F. Zhafran, E. S. Ningrum, M. N. Tamara, and E. Kusumawati, “Computer vision system based for personal protective equipment detection, by using convolutional neural network,” 2019.
- [18] M. I. B. Ahmed, L. Sarairoh, A. Rahman, S. Al-Qarawi, A. Mhran, J. Al-Jalaoud, D. Al-Mudaifer, F. Al-Haidar, D. AlKhulaifi, M. Youldash, and M. Gollapalli, “Personal protective equipment detection: A deep-learning-based sustainable approach,” *Sustainability 2023, Vol. 15, Page 13990*, vol. 15, p. 13990, 9 2023.

- [19] “Chvg conversion object detection dataset by scalersai.”
- [20] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, “Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset,” *Automation in Construction*, vol. 106, 2019.
- [21] L. Wang, L. Xie, P. Yang, Q. Deng, S. Du, and L. Xu, “Hardhat-wearing detection based on a lightweight convolutional neural network with multi-scale features and a top-down module,” *Sensors (Switzerland)*, vol. 20, 2020.
- [22] M. Zhao and M. Barati, “Substation safety awareness intelligent model: Fast personal protective equipment detection using gnn approach,” *IEEE Transactions on Industry Applications*, vol. 59, 2023.
- [23] N. Filatov, N. Maltseva, and A. Bakhshiev, “Development of hard hat wearing monitoring system using deep neural networks with high inference speed,” 2020.
- [24] R. C. Reyes, R. V. Sevilla, G. S. Zapanta, J. V. Merin, R. R. Maaliw, and A. F. Santiago, “Safety gear compliance detection using data augmentation-assisted transfer learning in construction work environment,” 2022.
- [25] N. D. Nath, A. H. Behzadan, and S. G. Paal, “Deep learning for site safety: Real-time detection of personal protective equipment,” *Automation in Construction*, vol. 112, 2020.
- [26] Q.-H. Tran, T.-L. Le, and S.-H. Hoang, “A fully automated vision-based system for real-time personal protective detection and monitoring,” *KICS Korea-Vietnam Int Jt Work Commun Inf Sci*, vol. 1, 2019.
- [27] F. Gong, X. Ji, W. Gong, X. Yuan, and C. Gong, “Deep learning based protective equipment detection on offshore drilling platform,” *Symmetry*, vol. 13, 2021.
- [28] N. Srija, S. Sridhar, S. Sudhir, M. V. Adithan, and G. Vignesh, “Safefaceyolo: Advanced workplace security through helmet detection and facial authorization,” *2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*, pp. 1–5, 3 2024.
- [29] S. Kumar, H. Gupta, D. Yadav, I. A. Ansari, and O. P. Verma, “Yolov4 algorithm for the real-time detection of fire and personal protective equipments at construction sites,” *Multimedia Tools and Applications*, vol. 81, 2022.
- [30] J. Zhou, Y. Zhang, T. Lu, and W. Fang, “Wearing detection by multi-scale information fusion network for live-line working scenarios,” 2022.
- [31] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, and Y. Zhao, “Fast personal protective equipment detection for real construction sites using deep learning approaches,” *Sensors*, vol. 21, 2021.
- [32] R. Bhana, H. Mahmoud, and M. Idrissi, “Smart industrial safety using computer vision,” *ICAC 2023 - 28th International Conference on Automation and Computing*, 2023.
- [33] C. Shi, D. Zhu, J. Shen, Y. Zheng, and C. Zhou, “Gbsg-yolov8n: A model for enhanced personal protective equipment detection in industrial environments,” *Electronics 2023, Vol. 12, Page 4628*, vol. 12, p. 4628, 11 2023.
- [34] M. Yakubraj, J. S. Rajkumar, S. S. Samuel, G. M. Palmer, and J. A. Jenifer, “Monitoring industrial protection gear using intelligent system,” *2024 2nd International Conference on Networking and Communications (ICNWC)*, pp. 1–6, 4 2024.
- [35] M. L. R. Collo, J. R. M. Esguerra, R. V. Sevilla, J. Merin, and D. C. Malunao, “A covid-19 safety monitoring system: Personal protective equipment (ppe) detection using deep learning,” 2022.
- [36] A. A. Protik, A. H. Rafi, and S. Siddique, “Real-time personal protective equipment (ppe) detection using yolov4 and tensorflow,” 2021.

- [37] M. L. Mokeddem, M. Belahcene, and S. Bourennane, "Real-time social distance monitoring and face mask detection based social-scaled-yolov4, deepsort and dsfd&mobilenetv2 for covid-19," *Multimedia Tools and Applications*, 2023.
- [38] S. Bashir, R. Qureshi, A. Shah, X. Fan, and T. Alam, "Yolov5-m: A deep neural network for medical object detection in real-time," 2023.
- [39] N. W. Bling and P. Sebastian, "Personal protective equipment detection with live camera," 2021.
- [40] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," 2018.
- [41] A. Sandru, G. E. Duta, M. I. Georgescu, and R. T. Ionescu, "Super - sam: Using the supervision signal from a pose estimator to train a spatial attention module for personal protective equipment recognition," 2021.
- [42] S. A. Karatopak and D. Sen, "Modular ppe detection system for workplace safety management using pose estimation," *14th International Conference on Electrical and Electronics Engineering, ELECO 2023 - Proceedings*, 2023.
- [43] R. Xiong and P. Tang, "Pose guided anchoring for detecting proper use of personal protective equipment," *Automation in Construction*, vol. 130, 2021.
- [44] A. M. Vukicevic, M. Djapan, V. Isailovic, D. Milasinovic, M. Savkovic, and P. Milosevic, "Generic compliance of industrial ppe by using deep learning techniques," *Safety Science*, vol. 148, 2022.
- [45] R. Varghese and S. M., "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6, 2024.
- [46] "Opencv - open computer vision library."
- [47] "Github - humansignal/labeling: Labeling is now part of the label studio community. the popular image annotation tool created by tzutalin is no longer actively being developed, but you can check out label studio, the open source data labeling tool for images, text, hypertext, audio, video and time-series data.."

APPENDIX A

Submitted Article

The article submitted to the journal Multimedia Tools and Applications is included in this appendix.

Detection of Personal Protection Equipment Based on Deep Learning

João Pedro Ferreira¹, Luís Nunes¹, Tomás Brandão¹

¹Instituto Universitário de Lisboa (ISCTE-IUL), Centro de Investigação
em Ciências da Informação, Tecnologias e Arquitetura, Avenida das
Forças Armadas, Lisbon, 1649-026, Lisbon, Portugal.

Contributing authors: jptfa@iscte-iul.pt; luís.nunes@iscte-iul.pt;
tomas.brandao@iscte-iul.pt;

Abstract

The detection of Personal Protection Equipment (PPE) plays a pivotal role in ensuring the well-being of workers, preventing serious injuries and fatalities (SIF). This paper aims to explore three distinct approaches using Deep Learning models to enhance the detection of PPE items such as hats, vests, gloves, boots, protective glasses, and ear muffs. The first approach involves training a unified model capable of detecting all PPE classes simultaneously across entire image contexts, providing a comprehensive view but facing challenges in detecting less prominent PPE items. The second approach adopts a sequential detection system: persons are initially detected using YOLOv8L, followed by identification of PPE items using a specialised model (PPE_all), achieving improved precision and recall. The third approach focuses on refining detection through Cropped Ground Truth Training, emphasising precise localisation and reducing background noise. Evaluation of these approaches using two test datasets and metrics of precision and recall highlights trade-offs between model complexity, precision, and recall. The findings offer insights into optimising PPE detection systems for real-world applications.

Keywords: Personal Protection Equipment (PPE) detection, Deep Learning, Workplace Safety, Computer Vision, Convolutional Neural Networks (CNNs)

1 Introduction

The correct use of Personal Protection Equipment (PPE) is instrumental in preventing Serious Injuries and Fatalities (SIF), playing a pivotal role in reducing the risks faced by workers. According to [1], there were 2.88 million non-fatal accidents related to work activities that resulted in four calendar days of absence in the European Union (EU). Additionally, approximately 3,300 accidents led to fatalities, highlighting the significance of effective PPE implementation in safeguarding the well-being of workers.

A study by [2] shed light on common missteps, including wearing the wrong type of PPE, failing to wear it altogether, or not wearing it correctly. These lapses can have dire consequences, ranging from cuts and burns to respiratory problems, and even death. To overcome this pervasive issue, employers must prioritise comprehensive training on PPE selection, use, and maintenance.

Real-time monitoring of PPE usage emerges as a proactive strategy to minimise the occurrence of these types of accidents. Numerous studies within this domain, as explored in this article, contribute with valuable insights into the development and implementation of effective PPE monitoring systems. These implementations can be broadly categorised into two main approaches: 1) *Traditional Approaches*, utilising methodologies such as Bluetooth Low Energy (BLE) or Radio-frequency Identification (RFID), and 2) *Computer Vision Deep Learning-based Approaches*, where Convolutional Neural Networks (CNNs) are employed in various forms, including Object Detectors and Pose Estimators.

In the upcoming sections, this paper will explore reality of workplace safety, introducing three innovative approaches to monitoring the correct use of PPEs in different situations. The methodologies presented include: (1) detecting persons and PPEs with the same object detector, (2) detecting persons and the PPEs at the same time (two different object detectors), and (3) detecting persons, cropping their bounding boxes, and then detecting the PPEs within the cropped images (running sequentially).

2 Background

As workplaces evolve, the need for advanced safety measures becomes increasingly crucial. The landscape of workplace safety is undergoing a transformation with the adoption of new technologies. Real-time Personal Protective Equipment (PPE) detection systems, leveraging innovations like Bluetooth Low Energy (BLE), Radio-frequency Identification (RFID), and cutting-edge Computer Vision Deep Learning-based approaches, are replacing conventional methods. This section lays the groundwork for an in-depth exploration of both traditional and deep learning approaches to PPE detection. Before delving into the methodologies, it is essential to understand the significance of PPE and its diverse types in ensuring a secure working environment. This subsection provides insights into the present and future of workplace safety regulations, emphasising the pivotal role of real-time PPE detection systems in safeguarding individuals in various occupational settings.

2.1 Personal Protective Equipment (PPE)

PPE refers to specialised gear or clothing designed to protect individuals from various workplace hazards, ensuring their safety and well-being. The types of PPE can vary based on the nature of the work environment and the specific risks involved. Figure 1 presents the most common types of PPE, such as:

- **Head Protection:** Helmets or hard hats safeguard against head injuries.
- **Eye and Face Protection:** Safety glasses, goggles, or face shields shield the eyes and face from debris, chemicals, or intense light.
- **Respiratory Protection:** Masks or respirators help filter out harmful airborne substances.
- **Hearing Protection:** Earplugs or earmuffs protect against excessive noise levels.
- **Hand and Arm Protection:** Gloves and arm coverings shield against cuts, burns, or chemical exposure.
- **Foot and Leg Protection:** Steel-toed boots or other specialised footwear safeguard against foot injuries.
- **Body Protection:** Coveralls, vests, or full-body suits provide overall protection against various hazards.



Fig. 1 Example of different types of PPE. (Adapted from [3])

Understanding the importance of proper PPE usage is crucial for maintaining a safe working environment. Real-time PPE detection systems play a vital role in ensuring compliance and addressing potential risks promptly.

2.2 Traditional Approaches

Traditional methods for enhancing workplace safety have leveraged proven technologies to address the challenges of real-time PPE detection.

Bluetooth Low Energy (BLE) has been used to mitigate the limitations of manual and visual PPE monitoring, which are labour-intensive and error-prone. For instance, [4] introduced a real-time PPE detection system using BLE tags and receivers to prevent the use of power tools without proper PPE, thereby reducing injury risks.

Radio-frequency Identification (RFID) technology has also been employed to automate PPE compliance monitoring. [5] developed a mobile RFID portal for construction sites, combining RFID tags, readers, and a user interface to provide real-time feedback and extensive data for safety analysis.

Operator Area Networks (OANs) offer localised communication networks between workers and their PPE. [6] proposed an OAN-based system using wearable tags and a centralised server to enhance real-time PPE compliance monitoring and generate data for targeted safety interventions.

The arrival of Smart Internet of Things (IoT) technologies provides another solution for automating PPE-tool pair checks in construction settings. [7] proposed a system integrating sensors, wireless networks, and a central monitoring system to offer real-time monitoring and data-driven insights for improved safety.

These traditional approaches have laid the groundwork for advancements in real-time PPE detection, highlighting their strengths and areas for improvement.

2.3 Computer Vision Deep Learning-based Approaches

Computer Vision Deep Learning (DL)-based approaches have emerged as a powerful tool for real-time PPE detection, offering enhanced accuracy, adaptability, and scalability. DL algorithms can be trained on vast datasets of images or videos to identify and localise PPE items with remarkable precision. This capability holds immense promise for enhancing workplace safety by enabling real-time monitoring of PPE usage and providing timely interventions to prevent potential hazards.

Object detection algorithms have significantly advanced with DL, enabling machines to perform such tasks with high accuracy and precision ([8]; [9]; [10]; [11]). DL models, particularly Convolutional Neural Networks (CNNs), have emerged as the backbone of modern object detection algorithms ([11]).

Object detection algorithms can be broadly categorised into two main types: two-stage and one-stage detectors ([9]; [10]). Two-stage detectors first generate object region proposals and then classify each proposal and refine its bounding box. Examples include R-CNN (Region-based Convolutional Neural Network) and Faster R-CNN ([11]). One-stage detectors directly predict bounding boxes and class probabilities for objects in a single stage, reducing computational cost. Examples include Single Shot MultiBox Detector (SSD) ([10]) and You Only Look Once (YOLO) ([9]).

Studies have demonstrated the efficacy of these approaches in PPE detection. For instance, [12] proposed a CNN architecture based on Faster R-CNN to identify and classify various PPE in an industrial setting, achieving high accuracy and precision. However, [13] showed that one-stage detectors could outperform this architecture in some scenarios.

One-stage detectors, particularly YOLO, have gained recognition across various domains. Implementations of the YOLO architecture for PPE detection have been

proposed by multiple authors [13–29], demonstrating great adaptability and impressive metrics.

The potential benefits of employing SSD are also noteworthy, given its speed advantage over YOLO. Studies by [30], [31], [32], and [33] have shown favourable outcomes using SSD, leaving room for exploring additional Artificial Neural Networks (ANN) architectures in future investigations.

Another groundbreaking application of DL is in pose estimation, enabling machines to accurately infer the positions and orientations of body parts from images and videos. Pose estimation algorithms can be classified into single-pose and multi-pose estimation. Single-pose estimation algorithms detect and track the positions of body joints for a single person, while multi-pose estimation algorithms handle multiple people, addressing challenges such as occlusions and variations in pose and appearance ([34]).

The integration of pose estimators into PPE detection has attracted significant attention. Studies by [35], [36], [37], and [38] have proposed innovative techniques combining pose estimation with object detection for accurate PPE recognition. These methods demonstrate the potential for enhanced accuracy in monitoring PPE usage in real-time.

In summary, the advancements in both traditional and DL-based approaches provide a robust foundation for developing effective real-time PPE detection systems. The integration of these technologies promises to significantly improve workplace safety by ensuring proper PPE usage and compliance.

3 Methodology

This section outlines the methodology employed to develop and evaluate three different systems for monitoring the correct use of PPE. Each system is designed with varying approaches to detect persons and their use of PPE, leveraging computer vision techniques and DL models. All models used in this study were trained using the 8th version of YOLO (YOLOv8), available at [39].

YOLOv8, the latest advancement in the YOLO series of object detection models, presented by [40], introduced significant improvements in both network architecture and performance. The structure of YOLOv8 is divided into two main components: the backbone network and the detection head.

The backbone network utilises EfficientNet-B4, a state-of-the-art neural network architecture. EfficientNet-B4, a 71-layer network with 19 million parameters, processes input images resized to 512x512 pixels and outputs five multi-scale feature maps. This allows YOLOv8 to extract detailed features from images, enhancing its detection capabilities.

The detection head in YOLOv8 is based on NAS-FPN, a neural architecture search method that creates feature pyramid networks for multi-scale predictions. YOLOv8 employs NAS-FPN-Cell, a sub-network that processes feature maps from the backbone network and generates bounding box predictions using anchor boxes to improve accuracy. Each feature map predicts three bounding boxes, resulting in a total of 15 bounding boxes for each input image.

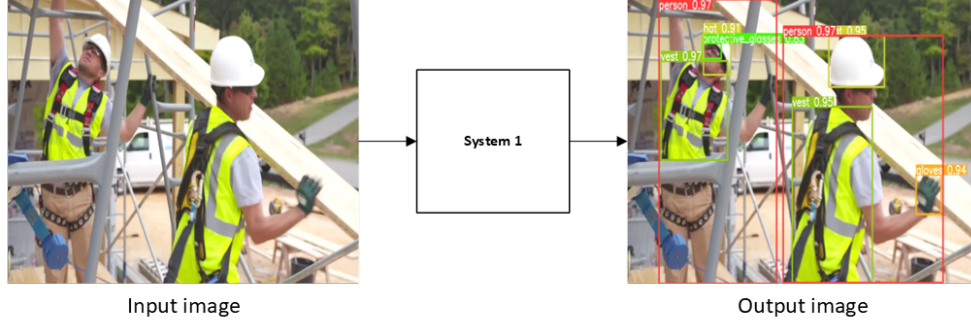


Fig. 2 Workflow of System 1 from input image to output with detected objects.

YOLOv8 also introduces several new features and enhancements that set it apart from previous versions. It adopts Focal Loss, a loss function that focuses on difficult examples and reduces the impact of easy ones, thereby improving recall and precision. For data augmentation, YOLOv8 uses Mixup, which blends two images and their labels to create new training examples, increasing data diversity and reducing overfitting. Additionally, YOLOv8 employs Average Precision Across Scales (APAS) as a new evaluation metric, measuring detection precision across different object scales and providing a more comprehensive performance assessment. These innovations make YOLOv8 a robust and efficient model for object detection, addressing many limitations of earlier YOLO versions.

3.1 System 1: Unified Detection

The first system utilises a unified approach, employing a single YOLOv8 model (specifically, the YOLOv8M variant) trained to detect both persons and various PPE items. This model is capable of identifying multiple objects within an image simultaneously, including persons, boots, ear muffs, gloves, hats, protective glasses, and vests. By training a single model to recognise all these objects concurrently, the system aims to make the detection process more efficient, offering a comprehensive solution that simplifies the overall workflow.

Figure 2 illustrates the workflow of System 1. The process begins with the input image, which is fed into the YOLOv8M model. The model then processes the image, performing object detection to identify and localise all relevant classes, including both persons and various PPE items. The output is an annotated image where each detected object is marked with bounding boxes and labels, indicating the type of object detected.

3.2 System 2: Sequential Detection with Two YOLOv8 Models

The second system adopts a sequential detection strategy using two distinct YOLOv8 models that process the entire image:

- **PPE Detection:** The cropped images are then fed into a YOLOv8 model (Cropped_PPE.all) trained to detect PPE items. By working on these smaller, person-specific images, the model can achieve higher accuracy in identifying PPEs such as boots, ear muffs, gloves, hats, protective glasses, and vests.

The cropping step in this system serves to reduce the complexity of the input for the PPE detection model, potentially improving detection performance by minimising the presence of irrelevant background elements. Figure 4 illustrates the workflow of System 3, from person detection to image cropping and final PPE detection.



Fig. 4 Overview of System 3: A two-stage detection process with image cropping for enhanced precision.

These three systems collectively offer a range of strategies for monitoring the correct use of PPEs using computer vision, each with its own advantages in terms of complexity, precision, and accuracy.

3.4 Data Preparation

For the data preparation phase, a selection of videos encountered in the internet was made to retrieve images for training and testing the systems. Four videos were chosen for this task, with two videos allocated only for training purposes and the remaining two only for testing.

3.4.1 Training and Validation Data

To gather training data, a simple script using OpenCV2 was developed. This script facilitated the conversion of video frames into individual images. Frames were sampled at intervals of 10 frames, resulting in approximately 3 images per second, ensuring a diverse representation of frames.

During the execution of the script, users interacted with it by pressing the 's' key. Upon pressing 's', the script initiated transforming video frames into images, capturing the frames at the specified intervals. This functionality provided users with control over the data collection process, allowing them to start and stop image capture as needed. Subsequent presses of the 's' key stopped the image capture process, providing users with flexibility in selecting frames for inclusion in the dataset. The idea behind this script was to ensure that only pictures with persons were saved and later used for model's training. The total number of images collected was 746, drawn from various segments of the videos. This sampling strategy ensured a comprehensive dataset encompassing a range of scenarios and conditions.

Following data collection, the dataset was partitioned into training and validation sets using a 70/30 split. This allocation was chosen to strike a balance between providing an wide amount of data for model training and retaining a separate set for validation. The 70% portion allocated to training facilitated the development of robust models, while the remaining 30% reserved for validation allowed for the evaluation of model performance on unseen data. This split was considered appropriate for preventing over-fitting and ensuring the generalisation of the trained models to unseen data.

Additionally, the Labellmg tool [41] was used to annotate the images by drawing bounding boxes around PPE items used by persons. PPE items not used by persons were not considered in this task.

For the evaluation of the trained models, it was used metrics such as Precision which measures the accuracy of the positive predictions made by the model, Recall which measures the ability of the model to correctly identify all relevant instances, Mean Average Precision at a confidence threshold of 0.5 (mAP50) and also Mean Average Precision across a range of confidence thresholds from 0.5 to 0.95 (mAP50-95).

3.4.2 Testing Data

For the testing phase, 50 random frames were selected from each video to avoid bias towards frames where the systems might perform better. Additionally, these frames were manually checked in order to guarantee that the frames were compliant with the type of images used to train the models.

To assess the system's performance across different scenarios, one video was selected that theoretically presented simpler conditions, making it easier for the system to predict correctly. Additionally, another video was chosen that presented more challenging conditions, posing greater difficulties for the system in accurately predicting Persons and PPE items.

Similar to the training data, the Labellmg tool was used to annotate the testing images by drawing bounding boxes around the PPE items.

This comprehensive data preparation process ensured that the systems were trained and tested on diverse and accurately annotated datasets, allowing for robust evaluation of their performance in detecting and identifying PPE items.

3.5 Training of Models

3.5.1 All Classes Detection (Person & PPE)

The initial phase of model training focused on training the YOLOv8 model to detect multiple classes, including persons and various PPE items such as boots, ear muffs, gloves, hats, protective glasses, and vests. The chosen model variant for this task was YOLOv8M. Training was conducted for 100 epochs to allow the model to learn patterns and features associated with each class. The model was trained to concurrently recognise and locate several objects inside an image by including many classes into the detection job. This approach offers a comprehensive solution for monitoring the appropriate use of PPEs.

Table 1 Results of YOLOv8 Model Training for Detection of All Classes

Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
All_classes	92.3	70.7	74.7	57.7

Table 1 summarises the performance metrics achieved by the YOLOv8M model trained for the detection of all classes. The Precision score of 92.3% indicates the accuracy of positive predictions made by the model, while the Recall score of 70.7% measures the model’s ability to correctly identify all relevant instances of the classes. The mAP50 scores of 74.7% and 57.7% for mAP50-95 reflect the model’s performance across different confidence thresholds, with higher scores indicating better performance. Overall, the YOLOv8M model demonstrates satisfactory performance in detecting all classes, achieving a balance between precision and recall.

3.5.2 Person detection

The initial phase of model training focused on training YOLOv8 models specifically for detecting persons. This step was crucial as it laid the foundation for subsequent training tasks involving the detection of Personal Protective Equipment (PPE) items.

To ensure accurate detection of persons, it was imperative to prepare the training data carefully. The training dataset consisted of images sampled from the collected dataset, ensuring a diverse representation of persons in various scenarios and conditions. These images were organised to exclude any bounding boxes corresponding to PPE items, thereby focusing solely on the detection of persons.

The training process involved training YOLOv8 models for 100 epochs, using two different variants: YOLOv8S and YOLOv8M. Each variant offered a different balance between model size, speed, and accuracy, allowing for experimentation to identify the optimal configuration for person detection.

The results demonstrate in Table 2 that both YOLOv8S and YOLOv8M models achieved high Precision and Recall scores, indicating their effectiveness in accurately detecting persons in the images. Additionally, the Mean Average Precision (mAP) scores suggest that both models perform well across a range of confidence thresholds, with YOLOv8M slightly outperforming YOLOv8S in terms of mAP50.

Table 2 Results of YOLOv8 Models Training for Person Detection

Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
YOLOv8S	98.2	96.5	98.0	86.9
YOLOv8M	98.4	96.4	97.6	87.5

Although the results were good, it was also evaluated the performance of YOLOv8 with the default weights in order to detect the class person. For that, it was used two variants of the model - YOLOv8M and YOLOv8L -, and it was also used the Ultralytics built-in task (validation).

Table 3 Results of YOLOv8 Models Validation for Person Detection

Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
YOLOv8M	96.2	88.3	95.1	82.9
YOLOv8L	96.3	85.5	95.8	84.6

Although the results of the trained models were better, it was decided to keep the default weights upon visual verification. This decision was made because the performance of the models with default weights, shown in Table 3 was considered sufficient for the task at hand, and using the default weights avoided the need for additional training time and computational resources.

To illustrate the performance differences between the trained YOLOv8 model and the default version, we present a visual comparison. In the Figure 5, the default YOLOv8 model successfully detects a person (highlighted in a red bounding box). However, the trained model fails to identify the person in the scene.

3.5.3 PPEs detection

In this part, three different approaches were employed to train object detectors for the detection of PPE. The goal was to enhance the model’s ability to accurately identify PPE items such as boots, ear muffs, gloves, hats, protective glasses, and vests. The following approaches were utilised:

Whole Image Training: In the first approach, the entire image was used for training the object detectors. This method allows the model to learn from the context of the whole scene, capturing both the person and the surrounding environment. Figure 6 shows an example of an image used in the training of such models.

Cropped Ground Truth Training: The second approach involved cropping the images to only include the regions corresponding to the ground truth bounding boxes of the persons. By focusing on these regions, the model was trained specifically on the areas where the PPE items are most likely to be found, potentially improving the precision of PPE detection. Figure 7 shows an example of an image used in the training of such models.

Masked Background Training: In the third approach, all areas outside the ground truth bounding boxes of the persons were masked out (painted black) in the original images. This technique emphasises the regions containing the persons while ignoring



Fig. 5 Comparison of person detection performance. The upper part of the image shows the detection result of the trained YOLOv8 model, which fails to identify the person. The lower part of the image shows the detection result of the default YOLOv8 version, which successfully detects the person (highlighted in a red bounding box).

the irrelevant background information. The objective was to improve the detection accuracy of PPE items on the subjects and minimise background noise by training the model on these masked photos. Figure 8 shows an example of an image used in the training of such models.

Each of these approaches was designed to explore different aspects of the training data, with the goal of improving the model’s ability to accurately detect and identify PPE items in various contexts. Additionally, the training of the models was divided into phases for each approach. Initially, the focus was on detecting the easier PPE items such as hats and vests. Subsequently, the models were trained to detect boots and gloves, and finally, the training included the more challenging items, protective glasses and ear muffs. This phased approach aimed to incrementally improve the model’s detection capabilities by starting with simpler tasks and progressively tackling more complex ones. For the training of such models, it was also considered the YOLOv8M version and the models were all trained for 100 epochs.

Starting off with the first approach, **1. Whole Image Training**, the models were trained progressively by introducing additional PPE classes. Initially, the focus was on training the model to detect simpler PPE classes such as hats and vests. Subsequently, the complexity of the models was increased by incorporating more classes, including boots and gloves, and finally including ear muffs and protective glasses as PPE classes.



Fig. 6 Example of image used in the training of the models



Fig. 7 Example of image used in the training of the models

For data preparation, the training data excluded the "Person" class along with its corresponding bounding boxes.

The results in Table 4 demonstrate that as the model's complexity increased, it maintained robust performance across different metrics. Notably, the PPE_hat_vest model achieved high precision (98.1%) and recall (94.0%), indicating its ability to accurately detect and classify hats and vests. As more classes were added (PPE_hat_vest_boots_gloves and PPE_all), there was a slight decrease in performance metrics, particularly in recall, suggesting a broader scope of detection challenges with additional PPE classes.

The performance decline in recall for the PPE_all model suggests challenges in detecting smaller or less distinct PPE items like protective glasses and ear muffs. This



Fig. 8 Example of image used in the training of the models

Table 4 Results of YOLOv8 Models Training for PPE Detection. Mean Average Precision at 50% IoU (mAP50) and Mean Average Precision at 50-95% IoU (mAP50-95) are metrics used to evaluate the accuracy of object detection models. mAP50 measures the precision at a single threshold (50% overlap), while mAP50-95 averages precision over multiple thresholds (from 50% to 95% overlap). Intersection over Union (IoU) measures the overlap between predicted and ground truth bounding boxes. (Values expressed in percentage)

Model	Precision	Recall	mAP50	mAP50-95
PPE_hat_vest	98.1	94.0	97.0	83.6
PPE_hat_vest_boots_gloves	90.8	78.9	84.9	64.7
PPE_all	91.7	65.3	69.7	52.3
Cropped_PPE_hat_vest	97.6	93.3	96.7	84.1
Cropped_PPE_hat_vest_boots_gloves	91.6	83.3	88.0	68.2
Cropped_PPE_all	93.2	83.8	88.8	60.9
Masked_PPE_hat_vest	97.7	93.2	96.6	83.6
Masked_PPE_hat_vest_boots_gloves	90.2	80.7	84.6	64.0
Masked_PPE_all	91.6	65.0	69.8	52.3

could be attributed to their variability in appearance and occlusion by other objects or body parts. However, the models consistently maintained high precision, indicating low false positive rates across all experiments.

Moving to the second approach, **2. Cropped Ground Truth Training**, significant adjustments were made to the dataset preparation process. In this method, each image suffered a transformation where the person of interest was cropped based on their ground truth bounding box. This approach was crucial to ensure that the training focused specifically on the regions containing persons and their associated PPEs.

To adapt the dataset for training under this approach, every image was processed to isolate the person of interest along with their PPE items. This involved extracting the relevant portion of the image defined by the ground truth bounding box of the person. Subsequently, all coordinates related to the PPE annotations were adjusted to align with the new cropped image dimensions.

The results in Table 4 demonstrate the effectiveness of training YOLOv8 models on cropped images. Notably, as the complexity of the model increases from detecting only hat and vest to including boots, gloves, and finally all specified PPE items, there is a trade-off observed between precision and recall. This trade-off is particularly evident in the broader category (Cropped_PPE_all), where the model achieves high precision but a slightly lower recall compared to the more focused categories. Overall, these results highlight the impact of dataset preparation on model performance, with cropped training data significantly enhancing the model’s ability to localise and identify PPE items accurately.

Moving to the third approach, **3. Masked Background Training**, the dataset preparation process involved a significant modification where all areas outside the ground truth bounding boxes of the persons were masked out (painted black) in the original images.

To prepare the dataset for training under this approach, each image experienced a transformation where only the portion inside the person’s ground truth bounding box was retained and the rest of the image was masked with black. This ensured that the model focused exclusively on the areas containing persons and their associated PPE items, without being distracted by the background context.

For the Masked Background Training approach, as shown in Table 4, the models consistently maintain high precision, indicating accurate identification of PPE items. However, there is a noticeable decrease in recall compared to the Cropped Ground Truth Training approach, particularly evident in the broader PPE_all category. This suggests that while masking the background helps in reducing noise, it may slightly affect the model’s ability to recall all instances of PPE items, especially in complex scenes.

4 Results and discussion

To evaluate the performance of the models, two different test datasets were used, each containing 50 images as mentioned previously. The metrics employed for evaluation were Precision and Recall. These metrics provide a comprehensive understanding of the model’s ability to accurately detect and classify PPE items.

For the evaluation of **System 1**, which involves the single-system detection, precision and recall were calculated by directly comparing the predicted bounding boxes to the ground truth. This straightforward approach allowed for an immediate assessment of the model’s accuracy in identifying PPE items.

However, for the other two systems, **System 2** and **System 3**, a more demanding evaluation criteria was applied. In these systems, a true positive was only considered valid if the PPE item was correctly predicted within a person who was also correctly predicted. This means that if a person was not predicted correctly but the PPE was, it was counted as a false negative. This evaluation method ensures that the models not only detect PPE items accurately but also associate them correctly with the detected persons, enhancing the reliability of the detection process.

4.1 System 1: Unified Detection

For the unified detection system, the model that was chosen is the **All_classes** model, which was discussed in the previous section. This model was selected due to its overall strong performance across various metrics, making it a reliable choice for detecting all PPE classes simultaneously.

Table 5 Performance of System 1 for each test dataset

Test Dataset	Precision (%)	Recall (%)
Test Data 1	73.0	45.7
Test Data 2	53.7	29.3

The results in Table 5 illustrate the performance of the All_classes model when applied to two different test datasets. For **Test Data 1**, the model achieved a precision of 73.0% and a recall of 45.7%. This indicates that while the model is reasonably accurate in identifying PPE items and persons, it still misses a significant number of true positives. The balance between precision and recall suggests that the model performs moderately well in this dataset but could benefit from further tuning to improve its recall.

For **Test Data 2**, the precision was 53.7%, indicating a lower accuracy in correctly identifying PPE items and persons compared to Test Data 1. The recall decreased further to 29.3%, highlighting a significant trade-off where the model is less precise and less capable of detecting all instances of PPE and persons. This could be due to variations in the dataset, such as different environmental conditions, lighting, or occlusions that make PPE and person detection more challenging.

Overall, the performance of the single-system detection using the All_classes model highlights the model’s strengths in precision but also underscores the need for further enhancements to improve recall. Future work could involve refining the model architecture, augmenting the training data, or incorporating additional pre-processing steps to address the challenges observed in recall.

4.2 System 2: Sequential Detection with Two YOLOv8 Models

In the second system configuration, the detection task is split into sequential steps: first detecting persons using YOLOv8L, which demonstrated superior performance in person detection in the previous sections, and subsequently detecting PPEs using the PPE_all model.

Following the successful detection of persons by YOLOv8L, the next step involves using the PPE_all model for detecting PPEs. This model was selected based on its performance across all PPE classes when trained on the entire image context. The system attempts to improve the memory and accuracy of recognising particular PPE items worn by observed humans by successively applying the PPE_all model after person detection.

Table 6 Performance of System 2 for each test dataset

Test Dataset	Precision (%)	Recall (%)
Test Data 1	86.4	83.5
Test Data 2	87.1	64.0

The results in Table 6 illustrate the performance of System 2, where person detection using YOLOv8L is followed by PPE detection using the PPE_all model, across two different test datasets.

For **Test Data 1**, the system achieved a precision of 86.4% and a recall of 83.5%. This indicates that the sequential approach of first detecting persons and then identifying PPE items using specialised models leads to high accuracy in both precision and recall. The high values in both metrics suggest that the system effectively detects and classifies PPE items on correctly identified persons.

Confusion Matrix

Actual	hat	70	0	0	0	0	0	10
	vest	0	63	0	0	0	0	14
	boots	0	0	5	0	0	0	11
	gloves	0	0	0	3	0	0	4
	ear_muff	0	0	0	0	0	0	0
	protective_glasses	0	0	0	0	0	0	2
	Background	2	2	10	0	0	0	0
		hat	vest	boots	gloves	ear_muff	protective_glasses	Background
		Predicted						

Fig. 9 Confusion Matrix for System 2 on Test Data 1

The confusion matrix in Figure 9 shows the results for System 2 on Test Data 1. The model correctly identified 70 instances of hats with only 2 false positives and it also identified 63 instances of vests with 2 false positives. However, there were challenges in detecting boots, where the model had 5 true positives but also 10 false positives. Gloves had 3 true positives and no false positives, indicating higher accuracy in this category. Ear muffs and protective glasses were not detected at all in this dataset.

For **Test Data 2**, the precision remained high at 87.1%, indicating consistent accuracy in identifying PPE items and persons. However, the recall decreased to 64.0%,

which suggests that while the system maintains high precision, it misses some instances of PPE items and persons present in the dataset.

		Confusion Matrix						
Actual	hat	36	0	0	0	0	0	24
	vest	0	19	0	0	0	0	8
	boots	0	0	9	0	0	0	22
	gloves	0	0	0	1	0	0	8
	ear_muff	0	0	0	0	0	0	0
	protective_glasses	0	0	0	0	0	5	4
	Background	3	0	2	1	0	0	0
		hat	vest	boots	gloves	ear_muff	protective_glasses	Background
		Predicted						

Fig. 10 Confusion Matrix for System 2 on Test Data 2

The confusion matrix in Figure 10 displays the results for System 2 on Test Data 2. Here the model correctly identified 36 instances of hats with 3 false positives and vests were identified correctly with 19 true positives and no false positives. Boots showed 9 true positives with 2 false positives and gloves had 1 true positive and 1 false positive. Ear muffs were not detected at all, while protective glasses had 5 true positives with no false positives.

Overall, the performance of System 2 using YOLOv8L for person detection and the PPE_all model for PPE detection demonstrates robust capabilities in accurately identifying PPE items on detected persons. The sequential approach appears effective in maintaining high precision, with scope for further enhancements to improve recall, particularly in challenging scenarios. Future work could involve optimising the model parameters, incorporating additional data augmentation techniques, or refining the training process to address the observed challenges in recall and further improve the system's overall performance.

4.3 System 3: Cropped Image Detection with YOLOv8

In the third system configuration, the detection task is approached in sequential stages: first detecting persons using YOLOv8L, which has demonstrated superior performance in person detection in the previous sections, and subsequently detecting PPEs using the Cropped_PPE_all model.

		Confusion Matrix						
Actual	hat	162	0	0	0	0	0	17
	vest	0	49	0	0	0	0	31
	boots	0	0	6	0	0	0	12
	gloves	0	0	0	1	0	0	6
	ear_muff	0	0	0	0	0	0	0
	protective_glasses	0	0	0	0	0	0	2
	Background	18	1	10	3	0	9	0
		hat	vest	boots	gloves	ear_muff	protective_glasses	Background
		Predicted						

Fig. 11 Confusion Matrix for System 3 on Test Data 1

Following the successful detection of persons by YOLOv8L, the next step involves using the Cropped_PPE_all model for detecting PPEs. This model was selected based on its performance across all PPE classes when trained on cropped images focusing on persons.

Table 7 Performance of System 3 for each test dataset

Test Dataset	Precision (%)	Recall (%)
Test Data 1	82.9	76.2
Test Data 2	84.2	62.4

The results in Table 7 illustrate the performance of System 3, where YOLOv8L is used for person detection followed by the Cropped_PPE_all model for PPE detection, across two different test datasets.

For **Test Data 1**, System 3 achieved a precision of 82.9% and a recall of 76.2%. This indicates that the sequential approach of first detecting persons and then focusing on the cropped regions for PPE detection leads to accurate identification of both persons and PPE items. The balanced precision and recall metrics suggest that the system effectively identifies and classifies PPE items on correctly identified persons.

The confusion matrix in Figure 11 shows the results for System 3 on Test Data 1. Notably, the model correctly identified 162 instances of hats with no false positives. Similarly, it identified 49 instances of vests with no false positives. Boots had 6 true positives with no false positives. Gloves had 1 true positive and no false positives. Ear muffs and protective glasses were not detected at all in this dataset. The background

Actual	hat	94	0	0	0	0	0	34
	vest	0	21	0	0	0	0	7
	boots	0	0	9	0	0	0	22
	gloves	0	0	0	2	0	0	7
	ear_muff	0	0	0	0	0	0	0
	protective_glasses	0	0	0	0	0	2	7
	Background	6	5	5	5	0	3	0
		hat	vest	boots	gloves	ear_muff	protective_glasses	Background
		Predicted						

Fig. 12 Confusion Matrix for System 3 on Test Data 2

column indicates the number of times the model incorrectly identified other objects as PPE items: 17 for hats, 31 for vests, 12 for boots, 6 for gloves, and 2 for protective glasses.

In **Test Data 2**, System 3 maintained a high precision of 84.2%, indicating consistent accuracy in identifying PPE items and persons. Nevertheless, the recall decreased to 62.4%, which suggests that while the system achieves high precision, it misses some instances of PPE items and persons present in the dataset.

The confusion matrix in Figure 12 displays the results for System 3 on Test Data 2. Here, the model correctly identified 94 instances of hats with no false positives. Vests were identified accurately with 21 true positives and no false positives. Boots showed 9 true positives with no false positives. Gloves had 2 true positives and no false positives. Ear muffs were not detected at all, while protective glasses had 2 true positives with no false positives. The background column indicates the number of times the model incorrectly identified other objects as PPE items: 34 for hats, 7 for vests, 22 for boots, 7 for gloves, and 7 for protective glasses.

Overall, the performance of System 3 using YOLOv8L for person detection and the Cropped_PPE_all model for PPE detection demonstrates robust capabilities in accurately detecting and classifying PPE items on identified persons using a sequential approach. The high precision values underscore the system's effectiveness in minimising false positives, while further refinements could focus on enhancing recall to capture more instances of PPE items and persons, especially in challenging scenarios. Future work could involve optimising the cropping strategy, fine-tuning the Cropped_PPE_all model parameters, or exploring additional techniques to address the observed challenges in recall and further elevate the system's overall performance.

5 Conclusion

In this study, we explored three distinct approaches for improving the detection of PPE using deep learning models. Each approach was designed to address different aspects of the detection task, ranging from holistic scene understanding to focused person-centric detection and precise PPE localisation.

The first approach involved training a single model, called **All_classes**, capable of simultaneously detecting all specified PPE items within the entire image context. This approach demonstrated strong precision, particularly for simpler PPE items such as hats and vests. However, as the complexity of PPE items increased, there was a noticeable trade-off with lower recall, indicating challenges in detecting all instances of PPE items across diverse scenarios.

In the second approach, we implemented a sequential detection system. Here, YOLOv8L was employed first to detect persons accurately, followed by the PPE_all model to identify PPE items. This sequential approach showed improved overall performance compared to the single-system approach, achieving higher precision and maintaining reasonable recall. This method leveraged the strengths of YOLOv8L in person detection and the thorough training of the PPE_all model on entire images, highlighting its effectiveness in identifying specific PPE items on correctly identified persons.

The third approach, focusing on Cropped Images Training, further refined the detection process. By training on cropped regions around persons, these methods aimed to enhance the model’s ability to precisely localise and classify PPE items directly associated with individuals. While this approach demonstrated competitive results in precision, maintaining low false positives, it showed a slight decrease in recall, particularly for more challenging PPE items such as protective glasses and ear muffs.

Overall, our findings underline the importance of tailored data pre-processing and model training strategies in improving the accuracy and reliability of PPE detection systems. Each approach demonstrated unique strengths and trade-offs, influencing the model’s performance across different evaluation metrics. Future research could focus on refining these approaches further, exploring ensemble methods, or integrating additional contextual information to enhance the robustness of PPE detection models in real-world settings.

Acknowledgment

This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) [ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020].

References

- [1] Eurostat: Accidents at work statistics (2023). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Accidents_at_work_statistics#Number_of_accidents

- [2] Sehsah, R., El-Gilany, A.H., Ibrahim, A.M.: Personal protective equipment (ppe) use and its relation to accidents among construction workers. *La Medicina del Lavoro* **111**, 285 (2020) <https://doi.org/10.23749/MDL.V111I4.9398>
- [3] SafetyToolBox: Personal Protective Equipment (PPE) (2023). <https://safetytoolbox.in/personal-protective-equipment-ppe/>
- [4] Gómez-de-Gabriel, J.M., Fernández-Madrigal, J.A., Carmen Rey-Merchán, M., López-Arquillos, A.: A safety system based on bluetooth low energy (ble) to prevent the misuse of personal protection equipment (ppe) in construction. *Safety Science* **158**, 105995 (2023) <https://doi.org/10.1016/J.SSCI.2022.105995>
- [5] Kelm, A., Laußat, L., Meins-Becker, A., Platz, D., Khazaei, M.J., Costin, A.M., Helmus, M., Teizer, J.: Mobile passive radio frequency identification (rfid) portal for automated and rapid control of personal protective equipment (ppe) on construction sites. *Automation in Construction* **36**, 38–52 (2013) <https://doi.org/10.1016/J.AUTCON.2013.08.009>
- [6] Pisu, A., Elia, N., Pompianu, L., Barchi, F., Acquaviva, A., Carta, S.: Enhancing workplace safety: A flexible approach for personal protective equipment monitoring. *Expert Systems with Applications* **238**, 122285 (2024) <https://doi.org/10.1016/J.ESWA.2023.122285>
- [7] Yang, X., Yu, Y., Shirowzhan, S., Sepasgozer, S., Li, H.: Automated ppe-tool pair check system for construction safety using smart iot. *Journal of Building Engineering* **32**, 101721 (2020) <https://doi.org/10.1016/J.JOBE.2020.101721>
- [8] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation (2014). <http://arxiv>.
- [9] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-December**, 779–788 (2015) <https://doi.org/10.1109/CVPR.2016.91>
- [10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9905 LNCS**, 21–37 (2016) https://doi.org/10.1007/978-3-319-46448-0_2/FIGURES/5
- [11] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* **28** (2015)
- [12] Zhafran, F., Ningrum, E.S., Tamara, M.N., Kusumawati, E.: Computer vision system based for personal protective equipment detection, by using convolutional

- neural network. (2019). <https://doi.org/10.1109/ELECSYM.2019.8901664>
- [13] Bashir, S., Qureshi, R., Shah, A., Fan, X., Alam, T.: Yolov5-m: A deep neural network for medical object detection in real-time. (2023). <https://doi.org/10.1109/ISIEA58478.2023.10212322>
 - [14] Zhou, J., Zhang, Y., Lu, T., Fang, W.: Wearing detection by multi-scale information fusion network for live-line working scenarios. (2022). <https://doi.org/10.1109/ICRCV55858.2022.9953241>
 - [15] Reyes, R.C., Sevilla, R.V., Zapanta, G.S., Merin, J.V., Maaliw, R.R., Santiago, A.F.: Safety gear compliance detection using data augmentation-assisted transfer learning in construction work environment. (2022). <https://doi.org/10.1109/CONECCT55679.2022.9865757>
 - [16] Kumar, S., Gupta, H., Yadav, D., Ansari, I.A., Verma, O.P.: Yolov4 algorithm for the real-time detection of fire and personal protective equipments at construction sites. *Multimedia Tools and Applications* **81** (2022) <https://doi.org/10.1007/s11042-021-11280-6>
 - [17] Nath, N.D., Behzadan, A.H., Paal, S.G.: Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction* **112** (2020) <https://doi.org/10.1016/j.autcon.2020.103085>
 - [18] Wang, Z., Wu, Y., Yang, L., Thirunavukarasu, A., Evison, C., Zhao, Y.: Fast personal protective equipment detection for real construction sites using deep learning approaches. *Sensors* **21** (2021) <https://doi.org/10.3390/s21103478>
 - [19] Tran, Q.-H., Le, T.-L., Hoang, S.-H.: A fully automated vision-based system for real-time personal protective detection and monitoring. *KICS Korea-Vietnam Int Jt Work Commun Inf Sci* **1** (2019)
 - [20] Gong, F., Ji, X., Gong, W., Yuan, X., Gong, C.: Deep learning based protective equipment detection on offshore drilling platform. *Symmetry* **13** (2021) <https://doi.org/10.3390/sym13060954>
 - [21] Collo, M.L.R., Esguerra, J.R.M., Sevilla, R.V., Merin, J., Malunao, D.C.: A covid-19 safety monitoring system: Personal protective equipment (ppe) detection using deep learning. (2022). <https://doi.org/10.1109/DASA54658.2022.9765088>
 - [22] Protik, A.A., Rafi, A.H., Siddique, S.: Real-time personal protective equipment (ppe) detection using yolov4 and tensorflow. (2021). <https://doi.org/10.1109/TENSYMP52854.2021.9550808>
 - [23] Bhing, N.W., Sebastian, P.: Personal protective equipment detection with live camera. (2021). <https://doi.org/10.1109/ICSIPA52582.2021.9576811>

- [24] Mokeddem, M.L., Belahcene, M., Bourennane, S.: Real-time social distance monitoring and face mask detection based social-scaled-yolov4, deepsort and dsfd&mobilenetv2 for covid-19. *Multimedia Tools and Applications* (2023) <https://doi.org/10.1007/s11042-023-16614-0>
- [25] Yakubraj, M., Rajkumar, J.S., Samuel, S.S., Palmer, G.M., Jenifer, J.A.: Monitoring industrial protection gear using intelligent system. 2024 2nd International Conference on Networking and Communications (ICNWC), 1–6 (2024) <https://doi.org/10.1109/ICNWC60771.2024.10537423>
- [26] Srija, N., Sridhar, S., Sudhir, S., Adithan, M.V., Vignesh, G.: Safefaceyolo: Advanced workplace security through helmet detection and facial authorization. 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), 1–5 (2024) <https://doi.org/10.1109/AIMLA59606.2024.10531484>
- [27] Bhana, R., Mahmoud, H., Idrissi, M.: Smart industrial safety using computer vision. ICAC 2023 - 28th International Conference on Automation and Computing (2023) <https://doi.org/10.1109/ICAC57885.2023.10275164>
- [28] Shi, C., Zhu, D., Shen, J., Zheng, Y., Zhou, C.: Gbsg-yolov8n: A model for enhanced personal protective equipment detection in industrial environments. *Electronics* 2023, Vol. 12, Page 4628 **12**, 4628 (2023) <https://doi.org/10.3390/ELECTRONICS12224628>
- [29] 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.: Personal protective equipment detection: A deep-learning-based sustainable approach. *Sustainability* 2023, Vol. 15, Page 13990 **15**, 13990 (2023) <https://doi.org/10.3390/SU151813990>
- [30] Filatov, N., Maltseva, N., Bakhshiev, A.: Development of hard hat wearing monitoring system using deep neural networks with high inference speed. (2020). <https://doi.org/10.1109/RusAutoCon49822.2020.9208155>
- [31] Zhao, M., Barati, M.: Substation safety awareness intelligent model: Fast personal protective equipment detection using gnn approach. *IEEE Transactions on Industry Applications* **59** (2023) <https://doi.org/10.1109/TIA.2023.3234515>
- [32] Wu, J., Cai, N., Chen, W., Wang, H., Wang, G.: Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. *Automation in Construction* **106** (2019) <https://doi.org/10.1016/j.autcon.2019.102894>
- [33] Wang, L., Xie, L., Yang, P., Deng, Q., Du, S., Xu, L.: Hardhat-wearing detection based on a lightweight convolutional neural network with multi-scale features and a top-down module. *Sensors (Switzerland)* **20** (2020) <https://doi.org/10.3390/>

- [34] Xiao, B., Wu, H., Wei, Y.: Simple Baselines for Human Pose Estimation and Tracking (2018). <https://github>.
- [35] Sandru, A., Duta, G.E., Georgescu, M.I., Ionescu, R.T.: Super - sam: Using the supervision signal from a pose estimator to train a spatial attention module for personal protective equipment recognition. (2021). <https://doi.org/10.1109/WACV48630.2021.00286>
- [36] Xiong, R., Tang, P.: Pose guided anchoring for detecting proper use of personal protective equipment. *Automation in Construction* **130** (2021) <https://doi.org/10.1016/j.autcon.2021.103828>
- [37] Vukicevic, A.M., Djapan, M., Isailovic, V., Milasinovic, D., Savkovic, M., Milosevic, P.: Generic compliance of industrial ppe by using deep learning techniques. *Safety Science* **148** (2022) <https://doi.org/10.1016/j.ssci.2021.105646>
- [38] Karatopak, S.A., Sen, D.: Modular ppe detection system for workplace safety management using pose estimation. 14th International Conference on Electrical and Electronics Engineering, ELECO 2023 - Proceedings (2023) <https://doi.org/10.1109/ELECO60389.2023.10416043>
- [39] Ultralytics: Home - Ultralytics YOLO Docs (2024). <https://docs.ultralytics.com/>
- [40] Varghese, R., M., S.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), pp. 1–6 (2024). <https://doi.org/10.1109/ADICS58448.2024.10533619>
- [41] GitHub - HumanSignal/labelImg: LabelImg is now part of the Label Studio community. The popular image annotation tool created by Tzutalin is no longer actively being developed, but you can check out Label Studio, the open source data labeling tool for images, text, hypertext, audio, video and time-series data. <https://github.com/HumanSignal/labelImg>