MIDOCEAN UNIVERSITY

FACULTY OF INFORMATICS

## Real-time Object detection of Personal Protective Equipment (PPE) in video Streaming Using YOLO Deep Learning Technique

BY

**Younis Arrabi  
Mohammad Al Johani**

~~A Thesis Submitted to~~

~~Department of Informatics in Partial Fulfillment of the Requirements for the Master Degree of Master of Artificial Intelligence~~

Under the Supervision of

### Prof. Dr. Yasser Abdelhamid

### Associate Professor, College of Informatics

### Midocean University

2024

## Acknowledgements

We want to express our gratitude to Prof. Dr. Yasser Abdelhamid for his guidance, encouragement and insightful feedback that played a significant role, in completing this research. His support has truly been a source of motivation.

Also extend thanks to the faculty and staff of the informatics college at Midocean University for creating an environment and providing the necessary resources for this research. Special appreciation goes out to my colleagues and friends whose support and engaging discussions have enhanced my research journey.

We are deeply grateful to our families for their support and encouragement. Their patience and understanding have been a source of strength, throughout this endeavor. Thank you all for your contributions and unwavering support.

Abstract

Magnificent progress in Artificial Intelligence computer vision techniques has been achieved during the last decades. Object detection should benefit from these improvements to discover about how workers are safe in different workplace environments. Because of the continuous increase in the number of researchers who are affiliated with computer vision and apply object detection techniques and deep learning methods, a relatively new discipline called Personal protective equipment auto detection has emerged. It aims to use methods, tools and techniques of deep learning in object detection to create and learn a model that can detect personal protective equipment in real-time and in automatic way.

Ensuring the proper use of Personal Protective Equipment (PPE) is critical in all workplace with accident high hazard settings; to reduce the risk, this project develops a real-time system for detecting Personal Protective Equipment in video streams using You Only Look Once (YOLO) deep learning techniques. The system identifies various Personal Protective Equipment requirements, such as helmets and vests, with high accuracy and minimal latency. By leveraging the efficiency of latest YOLO version, the solution achieves effective real-time performance, making it suitable for integration into existing surveillance infrastructures. This approach offers a scalable method for enhancing workplace safety monitoring and compliance.

Accuracy of detecting Personal Protective Equipment in most studies do not achieve the state-of-the-art accuracy in contrary with deep learning performances on datasets. The objective of this project is to develop a DL-based object detection model for the real-time detection of safety tools, including helmets, vests, gloves, safety shoes and safety goggles using different datasets.it will be designed to detect tools using object detection deep learning YOLO algorithm.

This can be decomposed into the following sub-objectives:

• Selecting and implementing proper algorithms from the recent literature for detecting objects.

• Evaluating the selected algorithms based on metrics such as accuracy, precision, and recall.

The research work introduced two deep learning models, namely Convolutional Neural Networks (CNNs), and You Only Look Once (YOLO), to detect PPE. The most used deep learning techniques in object detection are You Only Look Once (YOLO) and Convolutional Neural Networks so they must be included in our research. And their performance and advantage and limitation taken into considerations in selecting final Technique.

Our results showed that the YOLO models outperformed CNN Model. The accuracy of YOLO 98.4% and that of CNN was 98.2%.

The results show that accuracy of YOLO in Object detection reaches the state-of-the-art accuracies on datasets and can be efficiently used for detecting PPE of the workers and in real-time systems to reduce risk of accidents.

Contents

[Real-time Object detection of Personal Protective Equipment (PPE) in video Streaming Using YOLO Deep Learning Technique ii](#_Toc175087601)

[Prof. Dr. Yasser Abdelhamid ii](#_Toc175087602)

[Associate Professor, College of Informatics ii](#_Toc175087603)

[Midocean University ii](#_Toc175087604)

[Acknowledgements iv](#_Toc175087605)

[List of Abbreviations ix](#_Toc175087606)

[List of Figures ix](#_Toc175087607)

[List of Tables x](#_Toc175087608)

[Chapter 1 Introduction 11](#_Toc175087609)

[1.1 Problem Definition 13](#_Toc175087610)

[1.2 Thesis Objectives 15](#_Toc175087611)

[1.3 Thesis Contribution 16](#_Toc175087612)

[1.4 Thesis Organization 18](#_Toc175087613)

[Chapter 2 Literature Review 21](#_Toc175087614)

[2.1 Overview of PPE and Workplace Safety 21](#_Toc175087615)

[2.2 Sensor-Based PPE Detection Methods 22](#_Toc175087616)

[2.3 Computer Vision-Based PPE Detection Methods 28](#_Toc175087617)

[2.4 YOLO for PPE Detection 46](#_Toc175087618)

[Chapter 3 Related Work 47](#_Toc175087619)

[3.1 Detection of workers PPE 47](#_Toc175087620)

[3.2 The State-of-The-Art 68](#_Toc175087621)

[3.3 Summary 68](#_Toc175087622)

[Chapter 4 Methodology 68](#_Toc175087623)

[4.1 YOLO Algorithm 68](#_Toc175087624)

[4.2 Dataset Construction 72](#_Toc175087625)

[4.3 Training and Evaluation 77](#_Toc175087626)

[4.4 Performance Metrics and Benchmarking 80](#_Toc175087627)

[4.4 Summary 85](#_Toc175087628)

[Chapter 5 Experimental Results and Discussion 85](#_Toc175087629)

[5.1 Performance Evaluation 85](#_Toc175087630)

[5.2 Challenges and Limitations 86](#_Toc175087631)

[Chapter 6 Conclusion and Future Work 87](#_Toc175087632)

[7.1 Conclusion 87](#_Toc175087633)

[7.2 Future Work 88](#_Toc175087634)

[References 88](#_Toc175087635)

[[Abiodun, Oludare Isaac, et al., 2018] 88](#_Toc175087636)

[[Abu Saa, Amjed, et al., 2019] 89](#_Toc175087637)

[Appendices 89](#_Toc175087638)

# List of Abbreviations

* **PPE:** Personal Protective Equipment
* **ANN:** Artificial Neural Network
* **CNN:** Convolutional Neural Networks
* **R-**CNN: Region Based Convolutional Neural Networks
* **mAP:** Mean Average Precision
* **YOLO:** YOU ONLY LOOK ONCE
* **CM:** Confusion Matrix
* **DL:** Deep Learning
* **ML:** Machine Learning
* **AI:** Artificial intelligence

# List of Figures

Figure-1, RFID: The Technology Making Industries Smarter | TT Electronics

Figure-2, GPS Structure

Figure-3 Bluetooth beacon system

Figure 4 – Successful CNN

Figure 5 – R-CNN

Figure 6 - Picture Faster RCNN

Figure 7 - High-level diagram of SSD for generic object detection

Figure 8 – YOLO (Source Mastering All YOLO Models from YOLOv1 to YOLOv9: Papers Explained (2024) (learnopencv.com))

Figure 9 – YOLOv8 Architecture

Figure 10 - Annotation sample

Figure 11 – training Images

Figure 12 – validation Images

Figure 13 – Noise Augmentation

Figure 14 – Gray Scale Augmentation

Figure 15 – Rotation & Flipping Augmentation

Figure 16 – Scaling Augmentation

Figure 17 – mAP for YOLOv8 Object detection algorithm

Figure 18 – precession for YOLOv8 Object detection algorithm

Figure 19 – Recall for YOLOv8 Object detection algorithm

Figure 20 – Confusion Matrix

Figure 21 – All KPIs of the YOLOv8 Trained Model

# List of Tables

Table-2 Comparative summary for sensor-based PPE detection.

Table-3 Comparative summary between CNN & YOLO

Table 4 - Instances count

Table 5 – Showing instances and mAP & Recall metric.

Table 6 – F1 Score Results

## Chapter 1 Introduction

Safety of the worker during working hours become critical requirement over all kinds of projects, scaling from Small to mega size, companies and governments seeking to have safe workplace environment for the workers, and this have benefit on both workers and projects owners as well, these projects have complex tasks to maintain safety adherence and compliance, different project have different safety requirement, from complicated and simple workers-safety requirement, projects such as nuclear plants project, electricity main production stations, water desalination mega stations, chemical industries, firefighting, and also construction or small workshops.

In our projects we will focus on construction and industrial sectors that are experiencing rapid growth worldwide, making safety a primary concern for business owners and regulatory bodies.

Workers Safety can be achieved utilizing personal protective equipment (PPE) that can protect and avoid injuries and accidents to them, also saving project budgets and company reputations, Workers in these sectors face significant hazards that can lead to injuries and accidents.

To address these risks, numerous safety controls and protocols are implemented to ensure compliance with health and safety regulations across various workplace environments.

Several factors contribute to workplace accidents, including inadequate safety training, low awareness of safety measures, improper use of PPE, and insufficient presence of safety officers in hazardous areas.

Adhering to safety protocols and regulations can significantly reduce the risk of injuries, enhancing workplace safety and creating a better working environment. (OHSAS 18001 and ISO 45001)

Safety protocols typically include comprehensive training programs that cover safety rules, proper PPE usage, enforcement of PPE compliance, and periodic compliance checks.

The challenging in this task is that it is continuous and essential for protecting workers and minimizing the risk of injuries and accidents, mentioning key types of PPE are:

1. Head Protection: Helmets, hard hats, and bump caps to protect against head injuries.

2. Eye and Face Protection: Safety goggles, face shields, and glasses to protect against chemical splashes and flying particles.

3. Hearing Protection: Earplugs and earmuffs to reduce exposure to harmful noise levels.

4. Respiratory Protection: Masks and respirators to protect against inhaling hazardous substances.

5. Hand Protection: Gloves to protect against cuts, abrasions, chemicals, and extreme temperatures.

6. Body Protection: Protective clothing like lab coats, aprons, and coveralls to shield the body from chemicals, heat, and biological hazards.

7. Foot Protection: Safety shoes and boots to protect against injuries from heavy objects and electrical hazards.

8. Fall Protection: Harnesses and lanyards to prevent falls from heights.

Here, we will focus on PPE usage compliance by workers before entering the workplace, by detecting the PPE seamlessly utilizing different techniques, where the main goal of detection is real-time and accurate, this will be utilizing object detection multiple techniques.

[Jayaprakash et al. 2014].

### Problem Definition

The existing approach, to identifying breaches in using personal protective equipment involves manual checks and occasional audits. However, these methods fall short in safeguarding workers against accidents. Consequently, safety violations are often detected belatedly resulting in delayed interventions and heightened risks that could endanger the wellbeing of employees. The absence of a time automated mechanism, for overseeing and ensuring safety adherence poses an obstacle to the efficiency of established safety protocols.

The delay, in identifying and addressing safety breaches leads to an increase in workplace accidents, harm and deaths. These occurrences not impact the safety of workers. Also result in significant financial burdens for businesses, such, as healthcare costs, legal responsibilities and reduced efficiency. Moreover, recurrent safety issues can damage a company’s image. Invite rigorous regulatory oversight.

Advantages of Real Time Automated PPE Detection

1. Improved Safety; Real time detection systems can quickly spot PPE violations enabling actions to prevent accidents and injuries.
2. Cost Efficiency; Automated systems lessen the need, for checks and inspections resulting in substantial cost savings over time.
3. Compliance Boost; Continuous monitoring ensures adherence to safety rules enhancing compliance rates.
4. Reputation Management; Effective safety practices uphold a company’s reputation and minimize the risk of fines.

Challenges

1. Initial Capital Investment; Setting up real time detection systems involves an investment in technology and infrastructure.
2. Technical Hurdles; Ensuring the precision and dependability of automated systems in settings can pose challenges.
3. Privacy Concerns; The use of surveillance technologies raises worries about worker privacy and data protection.

The main objective is to create and implement a time automated system for monitoring PPE compliance. Such a system would improve worker safety cut costs and ensure adherence, to safety regulations ultimately fostering and more efficient work environment.

### 1.2 Thesis Objectives

The objective is to improve workers safety by creating a real time PPE detection system that can be utilized across various industrial environments ensuring adherence, to safety regulations and minimizing accident risks.

The main goal of this project is to create a learning model that can detect safety equipment, like helmets, vests, gloves, safety shoes and safety goggles in real-time by using different datasets. The model will be specifically designed to identify these tools using object detection algorithms.

This objective can be divided into the following sub goals;

1. Selecting and Implementing Algorithms; Apply the appropriate algorithms based on recent developments in object detection to identify safety equipment.
2. Evaluating Algorithms; Measure the performance of the chosen algorithms using metrics such as accuracy, precision and recall.

This project aims to build a real time system for detecting protective equipment (PPE) through deep learning methods. By utilizing algorithms, the system seeks to enhance safety compliance in settings.

### 1.3 Thesis Contribution

The main focus of this thesis is Testing a computer vision system that identifies Personal Protective Equipment (PPE) by utilizing the You Only Look Once (YOLO) algorithm. This study aims to improve safety adherence, in construction and industrial settings by applying cutting edge deep learning methods to enforce PPE regulations. The achievements of this research encompass aspects, such, as;

1. YOLO Model for PPE Detection: This thesis proposes the use YOLO to detect accuracy of PPE instances, addressing common challenges in real-world industrial settings.

2. Dataset Creation: A new dataset specifically tailored for PPE detection was created, comprising 1400 PPE instances. This dataset includes various types of PPE such as helmets, safety goggles, vests, gloves, and safety shoes, The dataset serves as a valuable resource for future research and development in this area.

1. Image Augmentation Techniques: The research introduces the use of the augmentation; Geometric Transformations, include Flipping, Rotation, Scaling, Translation. Also adding Noise to image and gray scale, that will enhance model detection results, pictures count reaches 9000 after applying the mentioned augmentation.

4. Performance Evaluation and Benchmarking: The YOLO model was evaluated, achieving a mAP (mean Average Precision) of 98.4%, maintained a real-time processing speed of from 1 to 22 FPS. These metrics highlight the model's effectiveness and suitability for real-time applications in industrial safety.

5. Comparative Analysis of Detection Methods: The thesis provides a comprehensive analysis of sensor-based and computer vision-based PPE detection methods. It highlights the limitations of traditional sensor-based approaches, such as high cost and operational challenges, and demonstrates the advantages of computer vision techniques in terms of accuracy, scalability, and cost-effectiveness.

6. Practical Implications and Future Work: The research outlines practical implications for deploying the proposed system in real-world industrial environments. It discusses the potential for integrating the system with existing safety protocols and provides recommendations for future work, including the exploration of other deep learning models and expanding the dataset to include more diverse scenarios and PPE types.

### References

1. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. \*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition\* (CVPR), 779-788.

2. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. \*arXiv preprint arXiv:2004.10934\*.

3. Glenn Jocher et al. (2021). YOLOv5 by Ultralytics. \*GitHub repository\*.

4. Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO Series in 2021. \*arXiv preprint arXiv:2107.08430\*.

5. Reza, A. M. (2004). Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. \*Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology\*, 38(1), 35-44.

### 1.4 Thesis Organization

The thesis is divided into chapters each focusing on elements of studying the identification of Personal Protective Equipment (PPE) with computer vision methods, particularly the YOLO algorithm. The layout guarantees a progression from laying out the background and purpose of the research to diving, into specifics, findings and conclusions.

1. Introduction

* Background and Motivation: Introduces the importance of PPE in ensuring workplace safety, the challenges associated with traditional detection methods, and the need for advanced computer vision solutions.
* Problem Statement: Clearly defines the research problem and the objectives of the study. (add objective)
* Contributions: Summarizes the main contributions of the thesis, including YOLO model and a new dataset for PPE detection.
* Thesis Organization: Provides an overview of the structure of the thesis.

2. Literature Review

* Introduction to Personal Protective Equipment (PPE) and Safety at Work; Examines the significance of using PPE in workplaces like construction settings.
* Detection Techniques for PPE Using Sensors; Explores conventional approaches to identifying PPE, such as RFID, infrared and ultrasonic sensors pointing out their constraints.
* Utilizing Computer Vision for PPE Detection; Looks into the advancements in computer vision methods for PPE detection with a focus on sophisticated algorithms, like CNNs and YOLO.

3. Methodology

* YOLO Algorithm; A thorough breakdown of the YOLO algorithm discussing its iterations and the progress made in each version.
* Building the Dataset; Describing how datasets are put together covering everything from gathering data to labeling and augmentation techniques.
* Training and Assessment; Exploring into the training procedures assessment measurements (mean Average Precision, Average Precision) and capabilities, for real time processing.

4. Experimental Results

* Assessing Performance; Discussing the outcomes achieved using the model while also comparing it to the initial YOLO.
* Result Examination; Digging into a review of how well the model performed, emphasizing enhancements in detecting personal protective equipment instances.
* Real life Scenarios; Showcasing instances, from situations where the suggested system was trialed showcasing its real-world usefulness.

5. Discussion

* Examining Traditional Approaches; Examines the pros and cons of using computer vision for detecting protective equipment in contrast to conventional sensor-based techniques.
* Obstacles and Boundaries; Points out hurdles and boundaries of the suggested method, including computing demands and applicability across various settings.
* Next Steps; Proposes areas for further study, like incorporating additional deep learning models and enlargement the dataset.

6. Conclusion

* Key Discoveries Summary; Summarizes the discoveries of the study highlighting their significance for improving workplace safety.
* Industry Implications; Explores how the research findings can practically enhance safety adherence in industrial and construction settings.
* Closing Thoughts; Offers reflections on the importance of the research and its potential for future progress, in the field.

7. References

- Comprehensive list of all sources cited throughout the thesis, formatted according to the appropriate academic style.

## Chapter 2 Literature Review

### 2.1 Overview of PPE and Workplace Safety

PPE usage compliance by workers before entering the workplace, by detecting the PPE seamlessly utilizing different techniques, where the main goal of detection is real-time and accurate, this will be utilizing object detection multiple techniques.

There are two main approaches to object detection they will enhance health and safety adherence, sensor-based and computer vision-based. Sensor-based techniques, which have been used since 2015, often involve high costs and additional safety risks. (https://doi.org/10.3390/su15010391).

Sensor-based PPE detection methods have been implemented in various industries to enhance safety compliance and reduce workplace accidents. These methods leverage different types of sensors to ensure that workers are wearing the required personal protective equipment (PPE) before entering or while present in hazardous areas., These sensor-based methods are critical in enhancing the safety of workers by ensuring they are equipped with the necessary PPE in real-time, although they come with challenges related to cost, practicality, and environmental constraints (<https://doi.org/10.3390/su15010391>)

### 2.2 Sensor-Based PPE Detection Methods

Going more depth, look at several sensor-based methods, Radio Frequency Identification (RFID) Tags, where Workers' PPE is equipped with RFID tags, which are detected by RFID readers installed at key locations such as entry points, exits, and within specific zones of a worksite, Advantages of RFID systems can track multiple items simultaneously and provide real-time monitoring. They are robust and can be integrated with existing safety management systems, saying about Limitations, the initial setup cost can be high due to the need for tags, readers, and integration infrastructure. There can also be issues with tag readability in environments with metal or liquid interference, we found application Example of it like on construction sites, RFID tags are embedded in helmets and vests. RFID readers at site entrances ensure that workers are wearing the correct PPE before allowing entry. Lee, S. J., & Park, M. (2010). Development of an RFID-based real-time safety management system on construction sites. Automation in Construction, 19(8), 1013-1021.

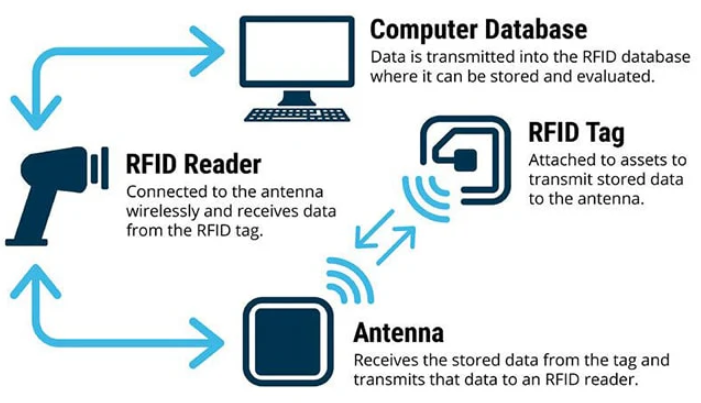


Figure-1, Photo courtesy of: [TT Electronics](https://blog.ttelectronics.com/rfid-technology), showing RFID system structure

[RFID: The Technology Making Industries Smarter | TT Electronics](https://www.ttelectronics.com/blog/rfid-technology/)

2nd sensor method, is Global Positioning System (GPS)-Based Systems, Workers carry GPS-enabled devices that track their location and ensure they remain within designated safe zones while wearing the required PPE, Advantages of GPS systems are effective in large, open environments and can provide detailed location data over extensive areas, mentioning the limitations of GPS-Based method, these systems are less effective indoors or in areas with poor satellite signal reception, GPS devices also require regular maintenance and power management, Application Example of it, In mining operations, GPS devices ensure workers are within safe zones and are equipped with necessary PPE such as respirators and helmets. Smith, R., & Jones, M. (2012). Using GPS to track and ensure compliance with PPE requirements. Journal of Safety Research, 43(5), 451-456.

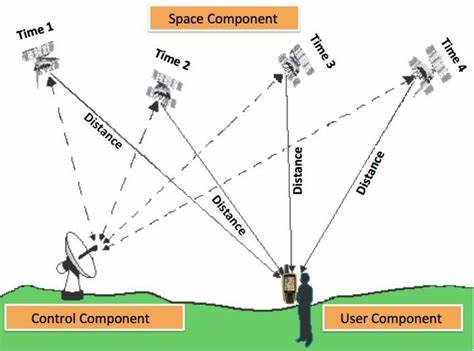


Figure-2 GPS Structure

Another method of sensor based in PPE detection is Bluetooth Beacons, PPE items are fitted with Bluetooth beacons that communicate with sensors located throughout the worksite, these sensors verify the presence of the beacons (and thus the PPE) in real-time, advantages of Bluetooth systems are relatively low-cost and suitable for indoor environments also they can provide precise location data within the worksite, mentioning its limitations, The range is limited compared to other technologies, and there can be interference from other Bluetooth devices or electronic equipment, can be used In manufacturing plants, Bluetooth beacons on ear protection and safety glasses communicate with sensors at workstations to ensure compliance before machine operation begins. Garcia, P., & Lopez, R. (2015). A Bluetooth-based monitoring system for PPE compliance. International Journal of Industrial Ergonomics, 50, 100-107.

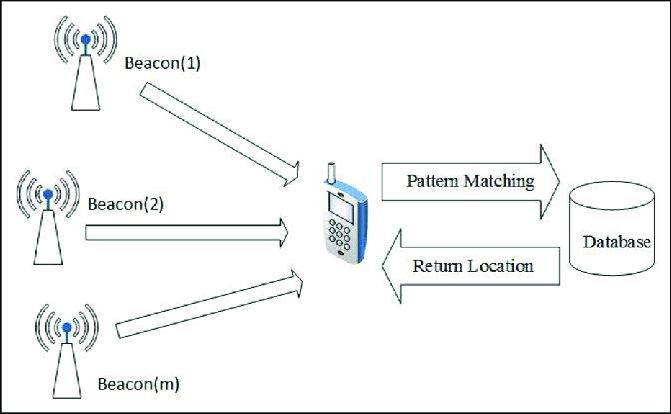


Figure-3 Bluetooth beacon system

4th method of sensor based in PPE detection is Proximity Sensors, it detect the presence of workers and their PPE by sensing specific tags or embedded sensors in the equipment, advantages of these sensors can provide immediate feedback and ensure that PPE is worn when workers approach hazardous equipment or areas, its limitations that Effective only in localized areas, requiring multiple sensors for comprehensive coverage, which can increase costs, application Example In chemical plants, proximity sensors near chemical handling areas detect if workers are wearing protective gloves and face shields. Johnson, T., & Martin, L. (2018). Proximity sensors for enhancing safety in industrial environments. IEEE Transactions on Industrial Informatics, 14(3), 1234-1242.

Last Method is Pressure Sensors, Pressure-sensitive mats or floors detect the presence of workers and can be linked to ensure PPE compliance when workers step on them, advantages is Provide direct and immediate feedback when workers enter restricted or hazardous areas, Limitations is high installation and maintenance costs and limited to specific entry or exit points, Application Example is Laboratories use pressure-sensitive mats at entrances to ensure workers wear lab coats and safety glasses before entering.

Kumar, N., & Singh, V. (2019). Pressure sensor applications in PPE compliance monitoring. Sensors and Actuators A: Physical, 285, 398-406.

Below is the table summary for sensor based different methods, which can give high level 360 view, and comparative summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method Name | Device | Advantages | Limitations | Application | References |
| Radio Frequency Identification (RFID) Tags | RFID TAG per PPE equipment  Antennas  RFID Readers Databases | real-time monitoring Integration with other monitoring system | High Cost,  Sensors is sensitive to Metal environment subject to interference | used at the site entrance & exits | Lee, S. J., & Park, M. (2010). Development of an RFID-based real-time safety management system on construction sites. Automation in Construction, 19(8), 1013-1021 |
| Global Positioning System (GPS)-Based Systems | GPS Devise Per PPE equipment  Database | effective in large, open environments and can provide detailed location data over extensive areas | Used in Outdoor High Operation Cost | mining operations, GPS installed in Helmet | Smith, R., & Jones, M. (2012). Using GPS to track and ensure compliance with PPE requirements. Journal of Safety Research, 43(5), 451-456. |
| Bluetooth Beacons | Bluetooth sensors per PPE Detectors | Low Cost Indoor Environment | Limited in Range Subject to interference | used in manufacturing plants | Garcia, P., & Lopez, R. (2015). A Bluetooth-based monitoring system for PPE compliance. International Journal of Industrial Ergonomics, 50, 100-107. |
| Proximity Sensors | Proximity sensors TAGs per PPE | Real-time | High operation cost | chemical plants | Johnson, T., & Martin, L. (2018). Proximity sensors for enhancing safety in industrial environments. IEEE Transactions on Industrial Informatics, 14(3), 1234-1242. |
| Pressure Sensors | Pressure-sensitive mats or floors detect the presence | Real-time | High Operation cost | Laboratories use pressure-sensitive mats at entrances to ensure workers wear lab coats and safety glasses before entering | Kumar, N., & Singh, V. (2019). Pressure sensor applications in PPE compliance monitoring. Sensors and Actuators A: Physical, 285, 398-406. |

Table-2 Comparative summary for sensor-based PPE detection.

Sensor-based PPE detection methods have played a significant role in enhancing workplace safety by ensuring compliance with safety protocols. These systems, utilizing different techniques as stated above.

Despite their benefits, such as precise tracking and alerting capabilities, they come with limitations including higher costs, potential discomfort for workers, and complex installation requirements.

### 2.3 Computer Vision-Based PPE Detection Methods

Building upon the groundwork established by sensor-based methods for detecting protective equipment (PPE) computer vision has emerged as a robust solution to enhance safety in the workplace.

By utilizing algorithms and deep learning methods computer vision systems can effectively identify and track the usage of PPE in real time.

In contrast to sensor-based approaches that rely on tags or devices computer vision employs cameras and image processing to recognize various types of PPE such as helmets, safety glasses and vests through object detection in the field of artificial intelligence.

Object detection involves the identification and localization of objects in images or videos typically represented by bounding boxes. This task has seen advancements over time with algorithms like Convolutional Neural Networks (CNNs) and YOLO (You Only Look Once) transforming object detection by providing real time processing capabilities and improved accuracy. This development has facilitated adaptable and scalable solutions across different domains like autonomous driving, surveillance and industrial automation. By harnessing datasets and powerful computational resources this technology offers benefits such as unobtrusive monitoring, scalability and operation, in diverse and complex environments.

With the advancements in deep learning models the integration of computer vision for Personal Protective Equipment (PPE) detection is poised to transform safety procedures enhancing compliance and reducing workplace injury risks significantly. Computer vision techniques utilize deep learning principles to recognize and ensure proper usage of PPE in real time. This approach typically involves utilizing cameras to capture images or videos, which are then analyzed using machine learning models to identify PPE items like helmets, safety glasses, vests and gloves.

Exploring deeper into these techniques Convolutional Neural Networks (CNNs) play a role. CNNs are a type of deep neural networks tailored for processing structured grid data such as images. They serve as a two-stage object detection algorithm widely employed for image recognition and classification tasks. Trained on labeled datasets comprising images of individuals with and without PPE CNNs learn to discern patterns and features associated with types of protective gear. This method offers advantages like accuracy and the capability to extract intricate patterns directly from the data; however, it necessitates sizable labeled datasets and substantial computational resources, for training purposes.

The CNN algorithm for object detection involves several key stages, as below

1. Input Layer
   1. **Image Acquisition**: The process begins with capturing or acquiring the image to be processed. The image is typically resized to a fixed dimension suitable for the CNN (e.g., 224x224, 416x416 pixels).
   2. **Normalization**: Pixel values are normalized to a range (e.g., [0, 1]) to facilitate faster convergence during training.
2. Convolutional Layers
   1. **Feature Extraction**: These layers apply convolution operations using various filters to detect features like edges, textures, and patterns. Each convolutional layer extracts higher-level features from the preceding layer.
   2. **Activation Function**: After convolution, an activation function is applied to introduce non-linearity to the model.
3. Pooling Layers
   1. **Down-sampling**: Pooling layers reduce the spatial dimensions of the feature maps, thus decreasing the computational load and controlling overfitting. Common pooling methods include max pooling and average pooling.
4. Fully Connected Layers
   1. **Classification**: These layers are used to classify the features extracted by the convolutional layers. They aggregate the features and predict the class probabilities.
   2. Dropout (Optional): Dropout layers are used to prevent overfitting by randomly setting a fraction of input units to zero during training.
5. Output Layer
   1. Final Predictions: The final output consists of the bounding boxes, class labels, and confidence scores for each detected object in the image.

Detailed Workflow in Object Detection with CNN:

1. **Preprocessing**: The image is preprocessed, normalized, and resized.
2. **Feature Extraction**: Convolutional layers extract hierarchical features from the image.
3. **Output**: The network outputs the final bounding boxes, class labels, and confidence scores.

References

1. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems.
2. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition.
3. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems.
4. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

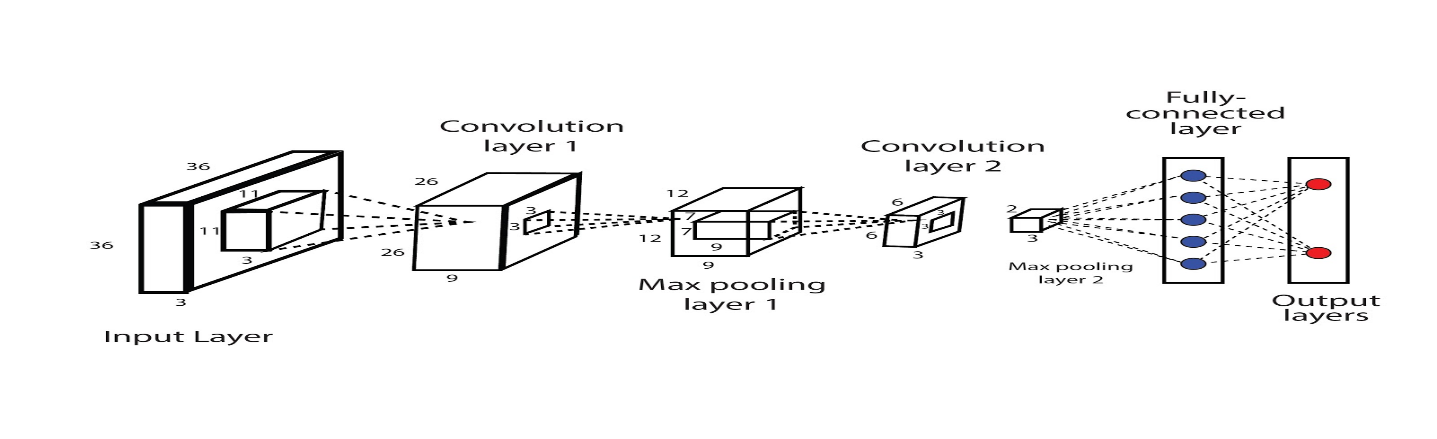


Figure 4 – Successful CNN

References: "Deep learning for safety surveillance: An approach to detect PPE compliance" (Chen et al., 2018).

2nd technique is Region-Based Convolutional Neural Networks (R-CNN), in R-CNN models divide the image into regions and then use CNNs to classify each region, it is Useful for detecting multiple PPE items in a single image, such as helmets, vests, and gloves, by focusing on specific regions of interest within the image, it has advantage that improved accuracy for object detection tasks by focusing on relevant regions, for the limitations it is Computationally intensive and slower than other methods.

The R-CNN algorithm for object detection involves several key stages, as below

1. Region Proposal Generation

**Selective Search**: This method generates a set of candidate object regions (also known as region proposals). It combines hierarchical grouping and greedy search to propose around 2,000 regions that are likely to contain objects.

1. Feature Extraction,

CNN Application: Each region proposal is warped to a fixed size and fed into a pre-trained Convolutional Neural Network to extract features. This step transforms the region into a fixed-length feature vector.

1. Classification

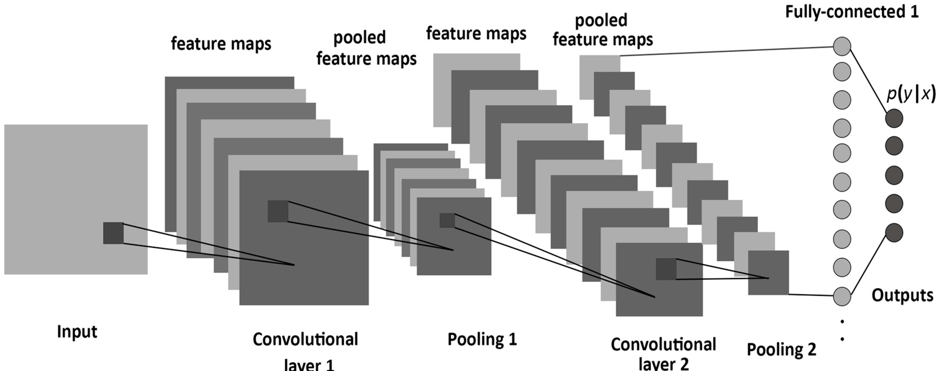
Classifiers: The extracted features from the CNN are fed into a set of Support Vector Machines (SVMs), one for each object class. The SVMs classify the regions into one of the pre-defined object categories or as background.

1. Output

Final Detections: The algorithm outputs the final set of bounding boxes, class labels, and confidence scores for the detected objects in the image.

References

1. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
2. Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective Search for Object Recognition. International Journal of Computer Vision, 104(2), 154-171.
3. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems.



Picture 5 – R-CNN

References: "R-CNN for real-time PPE detection in construction sites" (Girshick et al., 2014).

3rd is the Faster R-CNN, Method: An improved version of R-CNN, Faster R-CNN uses a Region Proposal Network (RPN) to generate candidate object bounding boxes and a subsequent network to classify these boxes, used for detecting PPE in more complex and cluttered environments by generating high-quality region proposals, advantages high accuracy and efficiency in generating and classifying region proposals, limitations for this technique more computationally expensive.

Here are the main stages of Faster R-CNN:

1. Convolutional Layers (Feature Extraction)

Convolutional Layers: The input image is passed through several convolutional layers (typically from a pre-trained network) to extract a rich hierarchy of features. These convolutional layers are shared between the Region Proposal Network (RPN) and the Fast R-CNN detector.

1. Region Proposal Network (RPN)

Anchor Boxes: The RPN slides a small network over the convolutional feature map and generates multiple region proposals (anchors) at each location. These anchor boxes are of different scales and aspect ratios.

Proposal Generation: The RPN outputs two scores for each anchor: an abjectness score indicating whether the anchor contains an object or not, and a bounding box regression to refine the anchor's coordinates.

NMS on Proposals: Non-Maximum Suppression (NMS) is applied to the proposals to reduce redundancy and keep only the most promising regions.

1. Region of Interest (RoI) Pooling

RoI Pooling Layer: The proposals generated by the RPN are projected onto the feature map. Each proposal is then warped into a fixed size using the RoI pooling layer, which extracts a fixed-length feature vector for each proposal.

1. Fully Connected Layers (Classification and Regression)

Object Classification: The feature vectors from RoI pooling are fed into fully connected layers to classify each proposal into one of the object classes or as background.

Bounding Box Regression: In parallel, the network refines the bounding box coordinates for each proposal to improve localization accuracy.

1. Output

Final Detections: The final output consists of the class labels and refined bounding boxes for each detected object. Non-Maximum Suppression is again applied to remove duplicate detections.

References

1. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems.
2. Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV).
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

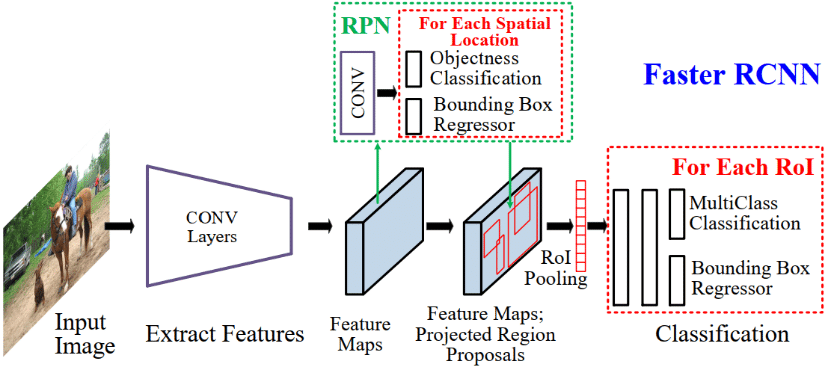


Figure 6 - Picture Faster RCNN

References: "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" (Ren et al., 2015).

4th Single Shot MultiBox Detector (SSD) is another technique, SSD detects objects in images using a single deep neural network that predicts both bounding boxes and class scores simultaneously, it is effective for detecting multiple PPE items in images with varying sizes and aspect ratios, for its Advantages, it Balances speed and accuracy, making it suitable for real-time applications, its Limitations that Generally, less accurate than Faster R-CNN but faster.

Here are the main stages of SSD:

1. Input Image and Base Network

Image Preprocessing: The input image is resized to a fixed size and normalized. This preprocessed image is then fed into a base network.

Base Network (Feature Extraction): A pre-trained convolutional neural network is used as the backbone to extract feature maps from the input image. These feature maps capture rich hierarchical features of the image.

1. Extra Feature Layers

Additional Convolutional Layers: SSD adds a series of auxiliary convolutional layers on top of the base network. These layers progressively decrease in size and act as feature detectors at multiple scales. They enable the detection of objects of various sizes.

1. Multi-Scale Feature Maps

Detection at Multiple Scales: SSD uses feature maps from different layers of the network (both from the base network and additional layers) to detect objects.

This approach allows the model to handle objects of various sizes by making predictions at multiple scales.

1. Default Boxes (Anchor Boxes)

Anchor Boxes: At each location on these feature maps, SSD places a set of default bounding boxes of different aspect ratios and sizes. These default boxes act as initial guesses for the presence and location of objects.

1. Classification and Localization

Convolutional Predictors: Each default box is evaluated by small convolutional filters that predict both the confidence scores for each class (including a background class) and the bounding box offsets. This is done for every location on each feature map.

Confidence Scores: These scores indicate the likelihood that a default box contains an object of a specific class.

Bounding Box Offsets: These offsets adjust the coordinates of the default boxes to better fit the actual objects.

1. Final Output

Detections: The final output consists of the refined bounding boxes, class labels, and confidence scores for each detected object in the image.

References

1. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. European Conference on Computer Vision (ECCV).
2. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., ... & Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
3. Liu, W., Wang, D., & Owens, J. D. (2021). Efficient convolutional neural networks for mobile devices. Journal of Parallel and Distributed Computing.

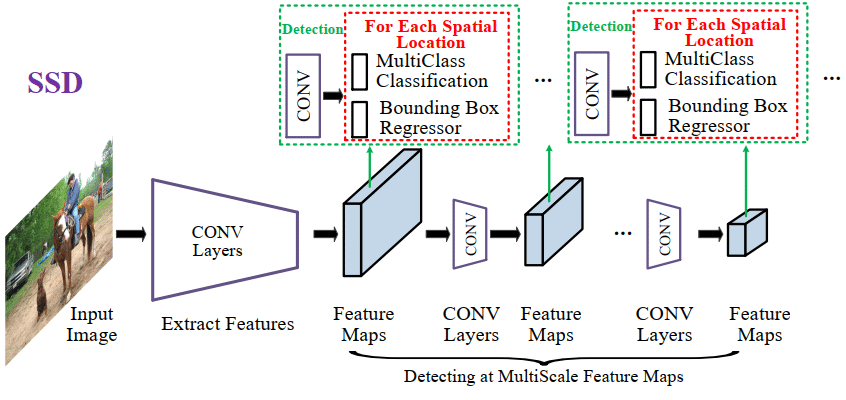


Figure 7 - High-level diagram of SSD for generic object detection

References: "SSD: Single Shot MultiBox Detector" (Liu et al., 2016).

Coming to most improved and latest technique in deep learning is You Only Look Once (YOLO), YOLO is an object detection algorithm that divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell in a single pass, it is ideal for real-time PPE detection due to its speed and accuracy, YOLO can simultaneously detect multiple types of PPE in real-time video streams, it is also have good advantages than others which is high speed and efficiency, suitable for real-time applications, for Limitations may struggle with detecting very small objects or objects in complex backgrounds.

The development of history began in 2016 and here are the key milestones in the evolution of YOLO;

YOLOv1 (2016); The initial release of YOLO (You Look Once) was pioneered by Joseph Redmon and his team. It transformed object detection by treating it as a regression issue directly forecasting bounding boxes and class probabilities from entire images in one pass. While YOLOv1 boasted speed, processing images at 45 frames per second it faced challenges with detecting small objects and precisely localizing objects within images.

YOLOv2 (2017); Building upon its predecessor, YOLOv2 introduced enhancements. It incorporated Batch Normalization, a High-Resolution Classifier and a novel anchor box method inspired by Faster R CNN. This version also utilized a revamped Darknet 19 architecture to improve both speed and accuracy. With the ability to detect over 9000 object categories.

YOLOv3 (2018); YOLOv3 marked another leap forward in model accuracy and capabilities. It embraced a network architecture called Darknet-53, which leveraged residual blocks and additional layers, for enhanced feature extraction. Introducing a scale detection approach enabled the network to identify small, medium and large objects more efficiently. Balancing speed with accuracy effectively positioned it for real time applications.

YOLOv4 (2020); This update aimed at enhancing performance and broadening the model’s accessibility to an audience. YOLOv4 introduced improvements, including using CSPDarknet53 as the foundation incorporating PANet for path aggregation and introducing a new data augmentation method called Mosaic. It achieved mean Average Precision (mAP) scores and quicker inference speeds compared to its predecessors.

YOLOv5 (2020); While not officially released by the YOLO creators YOLOv5 quickly became popular in the open-source community for its user-friendly nature and ease of implementation. YOLOv5 prioritized models, scalability and simplicity in both training and deployment processes. It provided model sizes (small, medium, large and extra-large) to strike a balance between speed and accuracy for different application scenarios.

YOLOv6 and YOLOv7 (2021 2022); These versions continued to enhance and optimize the architecture. Enhancements included efficient backbone structures improved training methodologies and enhanced multi scale prediction capabilities. Each iteration concentrated on pushing the boundaries of speed and accuracy establishing YOLO as a favored option for real time detection tasks.

YOLOv8 to YOLOv10 (2023 2024); The recent iterations have expanded on the strengths of prior versions by integrating advanced features, like Vision Transformers (ViTs) and more sophisticated augmentation techniques.

These new versions provide levels of precision and effectiveness making them perfect, for intricate object detection situations even in challenging lighting and environmental circumstances.

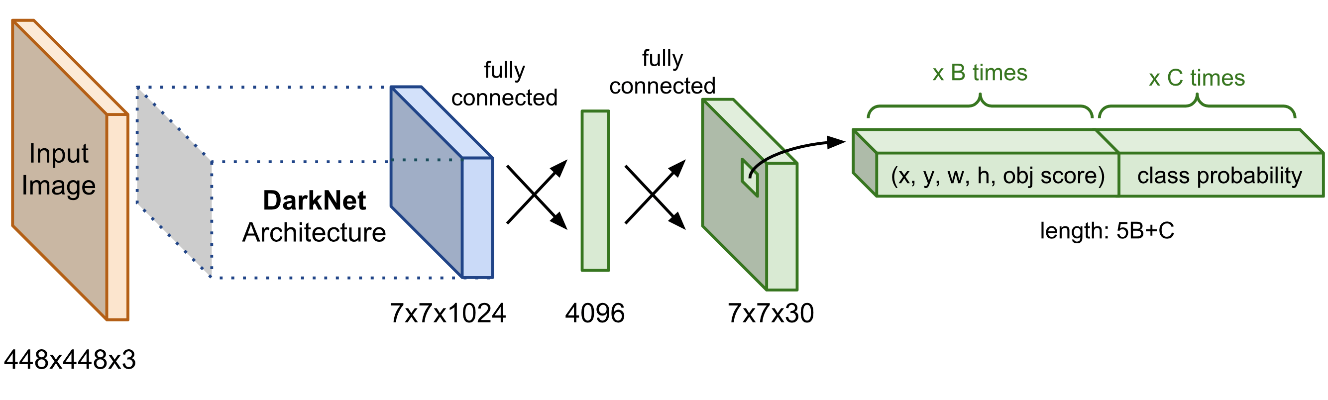


Figure 8 – YOLO (Source Mastering All YOLO Models from YOLOv1 to YOLOv9: Papers Explained (2024) (learnopencv.com))

References: "YOLOv3: An Incremental Improvement" (Redmon & Farhadi, 2018).

Detailed Comparison Between CNN and YOLO as below details, then we can select and have correct decision on best CV technique to detect PPE in real-time.

Convolutional Neural Networks (CNN)

Advantages:

1. High Accuracy: CNNs are known for their high accuracy in image classification and feature extraction tasks due to their ability to capture spatial hierarchies in images.

2. Flexibility: CNNs can be adapted for various tasks, including image classification, segmentation, and object detection, by modifying the network architecture and training strategy.

3. Deep Feature Learning: CNNs excel at learning deep hierarchical features from images, which is critical for understanding complex patterns and details.

4. Robustness to Variations: CNNs are generally robust to variations such as shifts, rotations, and scaling in the input images due to their use of convolutional filters and pooling layers.

Disadvantages:

1. Computational Cost: Training deep CNNs requires substantial computational resources and time, particularly for large datasets and complex models.

2. Overfitting: CNNs can easily overfit on training data if not properly regularized or if there is insufficient training data.

3. Complexity in Detection Tasks: While CNNs are excellent for classification, they require additional components (like region proposal networks in Faster R-CNN) to perform object detection, which can complicate the pipeline and increase computational demands.

You Only Look Once (YOLO)

Advantages:

1. Real-Time Detection: YOLO is designed for real-time object detection, processing images at high speeds frame per second (FPS), making it suitable for applications requiring instant feedback.

2. Single Pass Detection: Unlike other object detection methods that require multiple passes (like R-CNN variants), YOLO performs detection in a single pass, which simplifies the pipeline and reduces inference time.

3. Unified Architecture: YOLO frames object detection as a single regression problem, predicting bounding boxes and class probabilities directly from full images in a single evaluation.

4. Versatility: YOLO's design allows it to detect multiple objects of different sizes within the same image, and it performs well in various scenarios due to its multi-scale detection capabilities.

Disadvantages:

1. Localization Errors: YOLO can struggle with accurately localizing small objects within an image, particularly when there are multiple closely packed objects.

2. Lower Accuracy: Although YOLO is fast, it often has lower accuracy compared to other state-of-the-art detection models like Faster R-CNN in terms of mean Average Precision (mAP).

3. Fixed Grid Constraints: YOLO divides the image into a fixed grid and makes predictions based on these grid cells, which can lead to issues with detecting objects that span multiple cells or are located at the boundaries.

Must add references for below critical

|  |  |  |
| --- | --- | --- |
| **Metric** | **CNN-based Detectors** | **YOLO** |
| Accuracy (mAP) | Higher accuracy, better for complex scenes | Competitive accuracy, improving with each version |
| Speed (Inference Time) | Slower, not suitable for real-time applications | Fast, designed for real-time detection |
| Model Complexity | High complexity, requires more resources | Lower complexity, more efficient |
| Localization Accuracy | Very accurate in localizing objects | Improved localization, but still behind CNN-based models |
| Scalability | Less scalable without significant changes | Highly scalable with minor modifications |
| Training Time | Longer training times | Relatively efficient training |

Table-3 Comparative summary between CNN & YOLO

References

1. Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.
2. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934.
3. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems (NIPS).
4. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., ... & Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
5. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. European Conference on Computer Vision (ECCV).

### 2.4 YOLO for PPE Detection

YOLO (You Only Look Once) is a state-of-the-art object detection algorithm that frames detection as a single regression problem, predicting bounding boxes and class probabilities directly from full images in one evaluation. This methodology makes YOLO exceptionally fast and efficient compared to traditional region proposal-based methods like CNN and its variants.

Below are the advantages for using YOLO in PPE detection

1. Speed: The primary advantage of YOLO in PPE detection is its speed. Real-time detection is crucial for ensuring immediate feedback and enforcement of safety protocols on construction sites.

that YOLO can be deployed to continuously monitor environments for compliance with PPE usage, alerting supervisors instantly when violations occur.

3. High Throughput: YOLO's single-stage architecture allows for high-throughput processing, making it feasible to analyze video streams from multiple cameras simultaneously without significant latency.

4. Versatility: YOLO's multi-scale detection capability is advantageous for PPE detection, as it can identify various types of PPE (helmets, vests, gloves) within the same frame, ensuring comprehensive monitoring.

5. Implementation Simplicity: YOLO's unified architecture simplifies the implementation and deployment process, reducing the complexity compared to multi-stage detectors like Faster R-CNN.

## Chapter 3 Related Work

### 3.1 Detection of workers PPE

Chen and Demachi conducted a study that delved into a based technique for overseeing Personal Protective Equipment (PPE) in a nuclear power facility specifically honing in on identifying hard hats and full-face masks. The researchers made use of a labeled dataset comprising 3,808 images gathered from real world settings via webcams and web crawling. Their method involved training the YOLOv3 model in two phases using a combined dataset with a focus on freezing the final convolutional layer in Darknet 53 followed by fine tuning the entire network. The model exhibited performance metrics achieving precision of 97.64% and recall of 93.11% while maintaining a real time processing speed of 7.96 frames per second (FPS). The study’s key strength lies in its accuracy and utilization of real-world data although it does face limitations such as the relatively low FPS rate. Future endeavors will concentrate on enriching the dataset and integrating the model into on site monitoring systems to enhance accuracy further ultimately ensuring real time PPE identification, in hazardous environments.

Delhi et al. [8] leveraged deep learning techniques, particularly computer vision, to enhance safety on construction sites by enabling real-time PPE detection. The researchers manually collected a dataset of approximately 2,500 images, supplemented with images gathered through web scraping. The dataset focused on four target classes: NOT SAFE, SAFE, NoHardHat, and NoJacket. They employed YOLOv3, which was trained on augmented data, incorporating techniques such as flipping and 30-degree rotations to improve model robustness and generalization. The dataset was split into training, validation, and testing sets in a 90%, 8%, and 2% ratio, respectively. The model demonstrated strong performance, achieving a mean Average Precision (mAP), recall, and F1-score of 97% on the test data. The study's main advantage is its high accuracy in detecting safety compliance, though it is limited by the relatively small dataset size. The goal of the research was to develop a reliable system for real-time PPE detection, contributing to safer construction site environments.

In their study, Wang et al. [9] focused on enhancing worker safety through the application of deep learning neural networks tailored for real-time object detection to ensure adherence to safety protocols. They proposed three detectors based on the YOLO architecture—YOLOv3, YOLOv4, and YOLOv5—targeting six different classes. The research utilized a high-quality dataset known as CHV, comprising 1,330 images categorized into six classes, including four helmet colors, person, and vest. Among the models tested, YOLOv5x achieved the highest mean Average Precision (mAP) of 86.55%, while YOLOv5s provided the fastest performance with 52 frames per second (FPS). The study's advantage lies in its comprehensive evaluation of multiple YOLO variants, though it is limited by the dataset size. The primary goal was to develop an effective and efficient system for real-time PPE compliance monitoring.

Hayat and Morgado-Dias [11] introduced a deep learning approach for real-time detection of construction workers' heads and helmets. The researchers evaluated multiple versions of the YOLO architecture, specifically YOLOv3, YOLOv4, and YOLOv5x. They utilized the MakeML public dataset [12], comprising 3,000 instances for training and 1,000 for testing, focusing solely on the "Head" and "Helmet" classes. To address lighting and contrast issues in the images, they applied Power-law transformation [13] during pre-processing. Among the models tested, YOLOv5x delivered the best results, achieving 92% accuracy, 92.4% precision, 89.2% recall, and a 90.8% F1-score. The study highlights the effectiveness of YOLOv5x for accurate and efficient PPE detection, although it is somewhat limited by the narrow focus on just two classes. The goal was to enhance real-time safety monitoring on construction sites using advanced deep learning techniques.

Ma et al. [14] introduced a hybrid detection algorithm for PPE using the portable YOLOv4 model, applied to a dataset comprising approximately 25,000 samples obtained from security footage of a construction site. The dataset was divided into six uneven classes, and split into training and test sets. They utilized two algorithms, YOLOv4 and YOLOv4-Tiny, employing pruning as a fine-tuning and optimization strategy to enhance accuracy. The best results were achieved with CLSlim YOLOv4, which showed only a 2.1% mAP loss, increased inference speed by 1.8 times, and compressed model parameters by 98.2%. The study highlights the efficiency of the channel and layer pruning method (CLSlim) in reducing computational power usage and boosting detection speed. However, the study is limited by its focus on a single pruning method. Future research is suggested to combine CLSlim with other lightweight strategies to further accelerate model inference and explore better techniques for mobile devices with limited resources. Additionally, Gallo et al. [15] proposed a system for PPE detection in hazardous industrial areas using deep neural networks to analyze video streams. They trained five models—YOLOv4, YOLOv4-Tiny, SSD MobileNet, CenterNet, and EfficientDet—using three datasets, one of which was a publicly available set with 7,035 images, while the other two were collected in controlled settings, 215 and 236 images, respectively. YOLOv4-tiny, which achieved an mAP of 86%, was deployed in the system due to the speed of detection.

Nath, Behzadan, and Paal [17] proposed an AI-driven solution to address construction site fatalities, particularly those resulting from head injuries and collisions. Their study introduced a deep learning model based on the YOLO architecture to ensure PPE compliance. The process involves detecting workers and their PPE, such as hats and vests, and then using machine learning models like neural networks and decision trees to verify correct PPE usage. Another approach they explored utilizes a convolutional neural network (CNN) framework to simultaneously detect workers and confirm PPE compliance. A third method involves first identifying workers in the input image, followed by cropping and classifying them using CNN-based models like VGG-16, ResNet-50, and Xception to assess PPE presence. The models were trained on an in-house dataset called Pictor-v3, which contains 1,500 annotated images and 4,700 instances of workers wearing various PPE combinations [18]. The second approach yielded the best results in real-world environments, achieving a mean Average Precision (mAP) of 72.3% and real-time processing capability of 11 FPS, making it suitable for lightweight mobile devices. However, the study is limited by its relatively small dataset size and specific focus on certain PPE types. The primary goal was to enhance real-time safety monitoring on construction sites through advanced deep learning techniques.

Torrse et al. [19] introduced an innovative cognitive safety analysis component integrated into a monitoring system designed to ensure proper PPE usage in real-time. This system utilizes CCTV cameras to capture data, which is then analyzed by a deep learning algorithm for object detection. The study focused on the implementation of YOLOv4, achieving a mean Average Precision (mAP) of 80.19% at a real-time frame rate of 80 FPS. The key advantage of this approach is its high-speed processing capability, making it suitable for real-time safety monitoring. However, the study is limited by the specific focus on YOLOv4 and may benefit from exploring other detection models or hybrid approaches in future work. The primary goal was to enhance the real-time detection accuracy and efficiency of PPE compliance monitoring systems.

Isailovic et al. [22] focused on the detection of head-mounted protective gear by isolating the Region of Interest (ROI) around the head and testing several pre-trained computer vision models, including YOLOv5, MobileNetv2 SSD, and Faster R-CNN. The dataset used consisted of 12,682 images featuring 12 categories of PPE, sourced from the University of Belgrade’s Faculty of Medicine and public datasets like Roboflow [15] and Pictor PPE [17]. Among the models evaluated, YOLOv5 delivered the highest performance with 92% precision and 61.1% recall, particularly excelling in the "Hardhats" category, where it achieved 100% precision and 96% recall. The study highlights YOLOv5's superior accuracy, especially in detecting specific PPE categories, but acknowledges limitations in recall, suggesting potential improvements with more diverse training data or model enhancements. The goal was to advance real-time detection capabilities for PPE, emphasizing precision in safety-critical applications.

Cengil et al. [23] focused on enhancing the detection of hard helmets using a custom-improved one-stage object detector algorithm based on YOLOv5. The authors specifically improved the feature extraction process by experimenting with different backbone architectures—ShuffleNetv2 and MobileNetv3—to optimize the model’s efficiency and speed. Their study aimed to classify three objects: "Helmet," "Head," and "Person," using Roboflow’s "Hard Hat" dataset, which consists of 7,041 images split into training, testing, and validation sets. After evaluating the results, the YOLOv5 model with ShuffleNetv2 as the backbone achieved a higher precision of 94.2% compared to MobileNetv3, although both architectures delivered an equal recall rate of 91%. The study concluded that developing more comprehensive datasets with additional classes would be beneficial for future research, potentially enhancing the model's effectiveness in diverse real-world scenarios.

Kisaezehra et al. [24] explored enhancing safety at construction sites by employing a deep neural network, specifically YOLOv5, to detect whether workers are wearing safety helmets. The study utilized multiple network sizes ranging from nano to extra-large to optimize the balance between detection accuracy and processing speed. They utilized a dataset from Northeastern University in China, housed in Harvard Dataverse [25], comprising 7,063 images of construction workers in various locations and poses, both with and without helmets. The experiments demonstrated that the YOLOv5 extra-large model (YOLOv5x) delivered the best performance, achieving an mAP50 of 95.8%, a precision of 93.9%, a recall of 91.2%, and an F1-score of 92.5%. In contrast, the YOLOv5 nano model (YOLOv5n) was the fastest, showing superior speed in processing both images and videos. The study highlighted the potential for using varying model sizes to meet specific needs, suggesting further exploration into optimizing models for real-world applications.

The research by Lo, Lin, and Hung [26] aimed to develop deep learning models for real-time detection of Personal Protective Equipment (PPE) compliance using a dataset comprising 11,000 images. The study evaluated the performance of YOLOv3, YOLOv4, and YOLOv7, utilizing data preprocessing and augmentation techniques such as flipping, cropping, noise injection, and color space transformations to address the challenges of limited data and potential overfitting. The YOLOv7 model excelled, achieving the highest metrics with a mean Average Precision (mAP) of 97.29% and a processing speed of 25.02 frames per second (FPS). The study underscored the effectiveness of these models in enhancing workplace safety by ensuring PPE compliance, suggesting that future research could expand the dataset to cover more diverse conditions, improve detection of overlapping items, increase the number of negative training examples, enhance data collection methods, and identify additional necessary PPE types. It's important to note that despite the use of augmentation, the dataset predominantly featured images taken under conditions of high brightness and average weather, potentially limiting its general applicability in varying environmental conditions.

Lee, Yeo-Reum, and colleagues [27] developed a computer vision-based platform designed to monitor the correct usage of Personal Protective Equipment (PPE) on construction sites. They compiled a dataset consisting of 1,288 high-resolution images sourced from Google Images, surveillance cameras, and smartphones used at construction sites. This dataset, which includes categories such as Person, Hardhat, and Safety Vest, was divided into 1,031 training samples and 257 testing samples. The dataset facilitated the training of the YOLACT model, a state-of-the-art pixel-based PPE detection system. Additionally, the platform incorporated the DeepSORT algorithm for object tracking. The study found that YOLACT achieved a mean Average Precision (mAP50) of 66.4, and DeepSORT recorded a 91.3% accuracy rate in identifying whether workers were wearing PPE correctly. Future research aims to enhance the model’s capabilities to identify potential hazard situations by analyzing the interactions between construction site elements and the workers.

Ferdous and Ahsan [28] addressed the critical issue of worker safety at construction sites by implementing a computer vision approach. They trained the YOLOX-m model for the automatic detection of various PPE types, aiming to enhance protection levels. The study utilized a novel dataset named CHVG, comprising 1,699 images to teach the model to recognize eight classes, including four colored hardhats, vests, safety glasses, person body, and person head. To improve the robustness of the model, the dataset was augmented with geometric transformations (such as rotation, scaling, and translation) and photometric adjustments (such as brightness, HSV, and contrast changes). Additionally, manually written algorithms were used to generate images simulating challenging conditions like haze, rain, and low light. The YOLOX-m model outperformed other algorithms, achieving a mean Average Precision (mAP) of 89.84%. This approach's primary advantage lies in its high accuracy and ability to handle various environmental conditions, though it is limited by the dataset size.  
  
The goal was to develop a reliable system for real-time PPE detection to ensure worker safety in diverse and challenging environments.

Ke et al. [29] aimed to enhance safety by developing a deep learning-based real-time monitoring system for PPE compliance. The study utilized the YOLOv5 object detection algorithm to achieve high performance and speed. They trained the detection model using the FUZ-PPE dataset, which includes 18,767 annotated images of four types of PPE: helmets, masks, and safety wear. The dataset was split into 13,334 training images and 5,433 testing images. Optimization efforts reduced computational resources by 32% and training parameters by 25%. The resulting model achieved a detection speed of 105 FPS and a mean Average Precision (mAP) of 84.2%. Future enhancements suggested include expanding the FUZ-PPE dataset with images from real construction sites and improving the detection accuracy of the model. The study's main advantage is its high-speed performance and efficient resource usage, though it is somewhat limited by the initial dataset's scope. The goal was to create a robust, real-time PPE detection system to bolster construction site safety.

Saudi et al. [30] focused on ensuring construction worker safety by detecting three types of PPE: boots, hardhats, and vests. Workers wearing all three types were labeled "safe." The authors used a Region-based Convolutional Neural Network (R-CNN) architecture, known for its accuracy due to its two-stage detection process. They trained their model on data from the MIT database [31], which contains 1,129 construction site images, and evaluated it using 333 images collected personally by the authors. The model achieved a 70% accuracy rate. The authors plan to improve this accuracy by employing image processing techniques such as resizing and enhancing sharpness, as well as applying momentum optimization during training. The main advantage of this study is the high accuracy potential of the R-CNN architecture, though the current model's performance indicates room for improvement. The goal is to enhance the detection system to ensure better safety compliance on construction sites.

Mneymneh et al. [32] implemented a method to detect whether construction workers were wearing safety hardhats to prevent industrial site accidents using object detection techniques. They explored three different methods: feature extraction and matching, template matching, and cascade classification based on Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), and Haar-like features. The study involved training on 239 images from construction areas, with 75 images showing workers wearing hardhats and 164 without. Among the tested methods, the cascade classifier proved to be the most effective for various scenarios and real-time detection. The primary advantage of this approach is its adaptability and effectiveness in real-time applications, though it is limited by the relatively small dataset size. The goal was to develop a reliable detection system to enhance safety compliance and reduce accidents on construction sites.

Ji et al. [33] developed a model named RFA-YOLO, which integrates residual feature augmentation (RFA) with YOLOv4 to detect PPE worn by offshore drilling platform workers. The detection process consists of three stages: the first stage involves the RFA-YOLO model localizing and classifying individuals, helmets, and workwear. In the second stage, position features are constructed based on the detected bounding boxes. The third stage determines whether the identified helmet or workwear is being worn by the person. The model was trained and tested using the Offshore Drilling Platform Dataset (ODPD), which includes three subsets: the Object Detection Dataset (ODD) with 10,000 images across three classes, the Feature Classification Dataset (FCD) with 6,600 samples, and the PPE Dataset (PPED) with 2,000 images across four classes. The RFA-YOLO approach achieved an accuracy of 93.1% and a performance rate of 13 FPS. The main advantage of this method is its high accuracy in detecting PPE, although it is somewhat limited by its real-time processing speed. The goal was to enhance safety monitoring on offshore drilling platforms through accurate and efficient PPE detection.

Karlsson et al. [34] developed a system for facial recognition and PPE detection to ensure that only properly equipped individuals can access restricted areas. This camera-based system was tested for its ability to detect PPE at distances of 3 and 5 meters. The dataset used for training and evaluation was sourced from Kaggle [35]. The Faster R-CNN algorithm was employed for PPE detection, with data augmentation techniques such as adjustments in image brightness, contrast, blur, and sharpness to enhance model performance. Under controlled conditions, such as an airlock chamber, the system achieved a mean Average Precision (mAP) of 99% at 3 meters and 89% at 5 meters. The study concluded that the system could accurately identify PPE usage, particularly in collaborative environments. To address challenges with small objects or occlusions, techniques like data augmentation and body cropping were utilized. Future work suggests implementing the system at entry control points to verify PPE compliance before allowing access to restricted areas. The main advantage of this system is its high accuracy in controlled settings, though it may face limitations with occlusions and small item detection in more dynamic environments. The goal is to improve safety by ensuring proper PPE usage at critical access points.

Xiong and Tang [36] developed a framework for detecting PPE compliance using position-guided anchoring. They created the Personal Protective Equipment Dataset (CPPE), consisting of 932 images with 9,428 worker instances, to conduct their study. The framework operates by first employing a position detector to identify workers in a workplace setting, then using position-guided anchors to focus on specific body regions to detect PPE. Subsequently, two CNN-based classification algorithms were trained to determine whether these regions contained hardhats or vests. The proposed method achieved an F-score of 97% for hardhat detection and 95% for vest detection. However, the study faced limitations due to computational complexity, suggesting that the networks could be simplified to focus on detecting specific PPE items. The main advantage of this approach is its high accuracy in PPE detection, while its limitation lies in the computational demands. The goal was to enhance safety compliance by ensuring accurate and efficient detection of PPE in workplace environments.

Wang et al. [37] developed a computer vision-based method to enhance construction safety by tracking worker locations, checking for PPE usage, and predicting potential hazards. The dataset comprised images extracted from surveillance videos across various construction sites. The approach utilized an R-CNN model to track workers and their equipment, followed by a second CNN-based model to determine their movement paths. By analyzing the relationship between the detected gear and the workers, the system could identify potential threats and trigger alarms if necessary. The model achieved a mean Average Precision (mAP) of 92% and an Average Precision (AP) of 95% for detecting workers and gear, while the safety position analysis attained an 87% precision score. The primary advantage of this method is its high accuracy in monitoring and predicting safety issues, though it is limited by the computational complexity involved in processing surveillance video data. The goal was to proactively prevent safety incidents on construction sites through advanced real-time monitoring and hazard prediction.

Wang et al. [38] developed a detection method using a Convolutional Neural Network (CNN) to determine if construction workers are wearing hardhats, aiming to enhance safety by issuing alarms when violations are detected. The proposed method utilized MobileNet as the backbone for real-time detection, incorporating a top-down unit to improve feature extraction and a residual block to enhance classification accuracy. The dataset, created by the authors, includes images of construction workers both with and without hardhats. This approach outperformed other object detection models, achieving an Average Precision (AP) of 87% for instances without hardhats and 89% for instances with hardhats, while maintaining a detection speed of 62 frames per second (FPS). The main advantage of this method is its high detection accuracy and real-time performance. However, the study's limitation lies in its reliance on a specific dataset, which may affect generalizability. The goal was to implement an effective real-time monitoring system to improve worker safety on construction sites.

Li et al. [39] devised a real-time detection method for hardhats in construction areas to enhance safety by preventing work in hazardous environments. The dataset, comprising 3,261 images, was sourced from construction site surveillance systems and web crawlers. The SSD-MobileNet algorithm was employed for detection, achieving a precision of 95% and a recall of 77%. The study faced limitations due to the dataset's quality, as many images featured complex backgrounds and obscured hardhats, particularly for workers in power substations. The primary advantage of this approach is its high precision in detecting hardhats, although its recall was impacted by the challenging dataset. The goal was to implement a reliable real-time detection system to improve worker safety in construction environments.

Li et al. [40] introduced an advanced safety helmet detection framework utilizing computer vision, machine learning, and image-processing techniques aimed at creating an effective system for monitoring helmet usage. The dataset incorporated pedestrian data from power substations and the INRIA person dataset. The ViBe algorithm was employed for moving object segmentation, while the C4 pedestrian detection method and cascade classifier ensured accurate human detection. Analysis of ten videos demonstrated a mean pedestrian classification accuracy of approximately 84.2%. This proposed method surpassed existing techniques, achieving an area under the curve (AUC) of 94.13%, compared to 89.20% for methods using HOG feature extraction and an SVM classifier. The framework showed potential in accurately identifying safety helmet use through head location, color space transformation, and color feature differentiation. Future efforts will focus on enhancing the system’s robustness and applicability by extending its capability to handle more complex scenarios involving multiple pedestrians and challenging environments. The main advantage of this system is its high accuracy and advanced detection capabilities, though it currently faces limitations in handling highly complex scenes. The goal is to improve safety monitoring by ensuring reliable helmet detection in various environments.

Vukicevic et al. [41] proposed a solution for detecting PPE in industrial settings using several models, including MobileNetV2, DenseNet, and ResNet, all pre-trained on the ImageNet dataset. Although the models performed similarly, MobileNetV2 was identified as the most optimal due to its lower computational requirements. The dataset combined images from web mining and public PPE datasets. This approach is advantageous for its efficiency and low computational demand, though it is limited by the dataset's diversity. The goal was to create an effective and resource-efficient PPE detection system for industrial use.

Le and Si [42] developed a fully automated, vision-based system to enhance worker safety and reduce the risk of accidents. Their system detects PPE in the first stage and recognizes faces in the second stage. The primary aim of PPE detection is to ensure the presence of required safety gear, while face detection and recognition identify the workers. The system achieved up to 98% accuracy for PPE detection and up to 96% for face recognition. Real-time results demonstrated high precision and recall. Future work includes expanding the dataset to include equipment in different conditions and optimizing the code to increase detection speed. The main advantage is its high accuracy, but it is currently limited by the dataset scope. The goal is to provide a comprehensive safety monitoring system that ensures compliance and worker identification in real-time.

Maior et al. [43] developed a system for automatically detecting PPE in industrial sites using computer vision and machine learning. The dataset comprised 731 helmet images sourced from ImageNet, and the YOLOv2 algorithm was employed for detection. The study demonstrated the potential of using computer vision and machine learning to identify PPE usage through real-time video streaming, highlighting the importance of this technology in promoting a safer work environment and ensuring PPE compliance to prevent accidents. The study suggests expanding the model to detect various types of PPE simultaneously, enhancing the alarm system, and utilizing actual surveillance footage for more precise and accurate PPE monitoring in industrial settings. The main advantage of this approach is its ability to provide real-time monitoring, although it is currently limited by the narrow scope of the dataset and the types of PPE detected. The goal is to improve safety compliance and reduce accidents in industrial environments through advanced real-time detection systems.

Vibhuti et al. [44] developed an automated method using a transfer learning model to detect individuals not wearing masks in public settings during the COVID-19 pandemic. The study employed several deep learning models, including InceptionV3, Xception, MobileNet, MobileNetV2, VGG16, and ResNet50, and used the Simulated Masked Face Dataset (SMFD) [45] for training and testing. The InceptionV3 model was optimized through fine-tuning. The results showed the method achieved 100% accuracy and specificity during testing and 99.92% accuracy during training. This highlights the effectiveness of the proposed transfer learning model in accurately identifying non-mask wearers. The study suggests that future work should involve using larger datasets, expanding the system to classify different types of masks, and integrating a facial recognition system to identify individuals while wearing masks. The key advantage is its high accuracy in detecting mask compliance, while the limitation lies in the need for larger datasets and more varied mask types. The goal is to enhance public health safety by ensuring mask compliance through advanced detection technology.

summary of literature review including the techniques being used along with the type and source of dataset and major achievements of the studies.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Link** | **Year** | **Goal** | **Algorithm** | **Dataset** | **Performance measure** |
| 6 | 2020 | Vision-based approach for monitoring PPE in a Nuclear power station | YOLOv3 | 3808 images collected from the web | Precision of 97.64% and a Recall of 93.11% |
| 7 | 2020 | Ensure a maximum level of safety at the construction sites by detecting the PPE in a real-time |  | 2500 images collected manually | Precision, Recall, and F1-score of 97% |
| 8 | 2021 | Real-time object detection to ensure workers’ compliance with safety measures | YOLOv3, YOLOv4, and YOLOv5 |  | YOLOv5x: best mAP 86.55% |
| 10 | 2022 | Detect construction site workers’ heads and helmets in real-time | YOLOv3, YOLOv4, and YOLOv5x |  | YOLOv5x with accuracy of 92%, precision of 92.4%, recall of 89.2%, and F1-score of 90.8% |
| 13 | 2022 | Detect construction site workers’ PPE | YOLOv4 and YOLOv4-Tiny | 25,000 samples taken from security footage of a building construction site | CLSlim YOLOv4 mAP loss of 2.1% |
| 14 | 2022 | Detect PPE in unsafe industrial areas | YOLOv4, YOLOv4-Tiny, SSD MobileNet, CenterNet, and EfficientDet | 15 | YOLOv4-tiny, mAP of 86% |
| 16 | 2020 | Intelligence-based solutions to resolve construction fatalities caused by brain injuries and collisions | YOLO and CNN | Pictor-v3 [17] | CNN 72.3% mAP |
| 18 | 2021 | New cognitive safety analysis component for a monitoring system | YOLOv4 | CCTV | mAP of 80.19% |
| 21 | 2022 | Detection of head-mounted protection gear | YOLOv5, MobileNetv2 SSD, and Faster R-CNN |  | YOLOv5 with 92% precision and 61.1% recall |
| 23 | 2023 | Detect if the worker is wearing a hat or not | YOLOv5 (nano, Small, medium, large, and extra-large) | 24 | YOLOv5x mAP50 of 95.8%, precision of 93.9%, recall of 91.2%, and F1-score of 92.5% |
| 25 | 2022 | Create deep learning algorithms for the real-time detection of PPE | YOLOv3, YOLOv4, and YOLOv7 | 11000 | YOLOv7 with mAP value of 97.5% |
| 27 | 2022 | Providing the highest level of protection | YOLOX-m | CHVG | mAP of 89.84% |
| 26 | 2023 | Monitor the proper use of PPE on construction sites | YOLACT. Besides, DeepSORT. | Person, Hardhat, and Safety vest | DeepSORT accuracy 91.3% |
| 28 | 2022 | Improving safety by introducing a deep learning real-time monitoring system for the PPE | YOLOv5 | FUZ-PPE | 105 FPS and an mAP of 84.2% |
| 29 | 2020 | Ensure the safety conditions of construction workers | R-CNN | 30 | 70% |
| 31 | 2020 | Detecting whether construction workers are wearing safety hardhats or not to prevent accidents at industrial sites | Classification method based on HOG |  |  |
| 32 | 2023 | Detecting the PPE of offshore drilling platform workers | YOLO | ODPD | Accuracy of 93.1% and performance of 13 FPS |
| 33 | 2023 | Develop a facial recognition and PPE-detection system that could be applied at entry points to restricted locations | Faster R-CNN | 34 | mAP of 99% at 3 m and 89% at 5 m |
| 37 | 2020 | Determine if workers are wearing hardhats or not and to alarm them by using a CNN | CNN | 24 | AP of 87% for the hardhat negative instances and an AP of 89% for the positive instances, detecting within 62 FPS |

Table 4 – Summary for related works

### 3.2 The State-of-The-Art

### 3.3 Summary

## Chapter 4 Methodology

### 4.1 YOLO Algorithm

The YOLO algorithm, recognized for its ability to detect objects in real-time is chosen for its effectiveness and precision. This section explores into an overview of the YOLO framework charting its progress, as the primary objective of this research is to develop a real-time system for detecting Personal Protective Equipment (PPE) in various industrial settings using the YOLOv8 (You Only Look Once, Version 8) algorithm. YOLOv8 represents one of the latest advancements in object detection technology, offering improvements in speed, accuracy, and efficiency. This methodology outlines the steps involved in preparing the dataset, training the YOLOv8 model, evaluating its performance, and implementing it for real-time PPE detection.

Utilizing YOLOv8 for PPE detection involves connecting its object detection capabilities to improve workplace safety by automatically recognizing the presence and correct usage of personal protective equipment (PPE) in real time. YOLOv8 as a cutting-edge deep learning model delivers precision and speed making it effective for identifying various PPE items like helmets, gloves and safety vests in dynamic and intricate environments. The implementation of YOLOv8 for PPE detection can contribute to reducing workplace incidents ensuring adherence to safety protocols and offering feedback to employees and supervisors, in industrial settings.  
YOLOv8 Model Components is consist of 3 main blocks, the backbone is responsible for extracting features from input images. It consists of convolutional layers followed by activation functions, neck is the second block which aggregates features from different layers to make the model robust to objects of various sizes, last one is the head, it predicts bounding boxes and class probabilities.

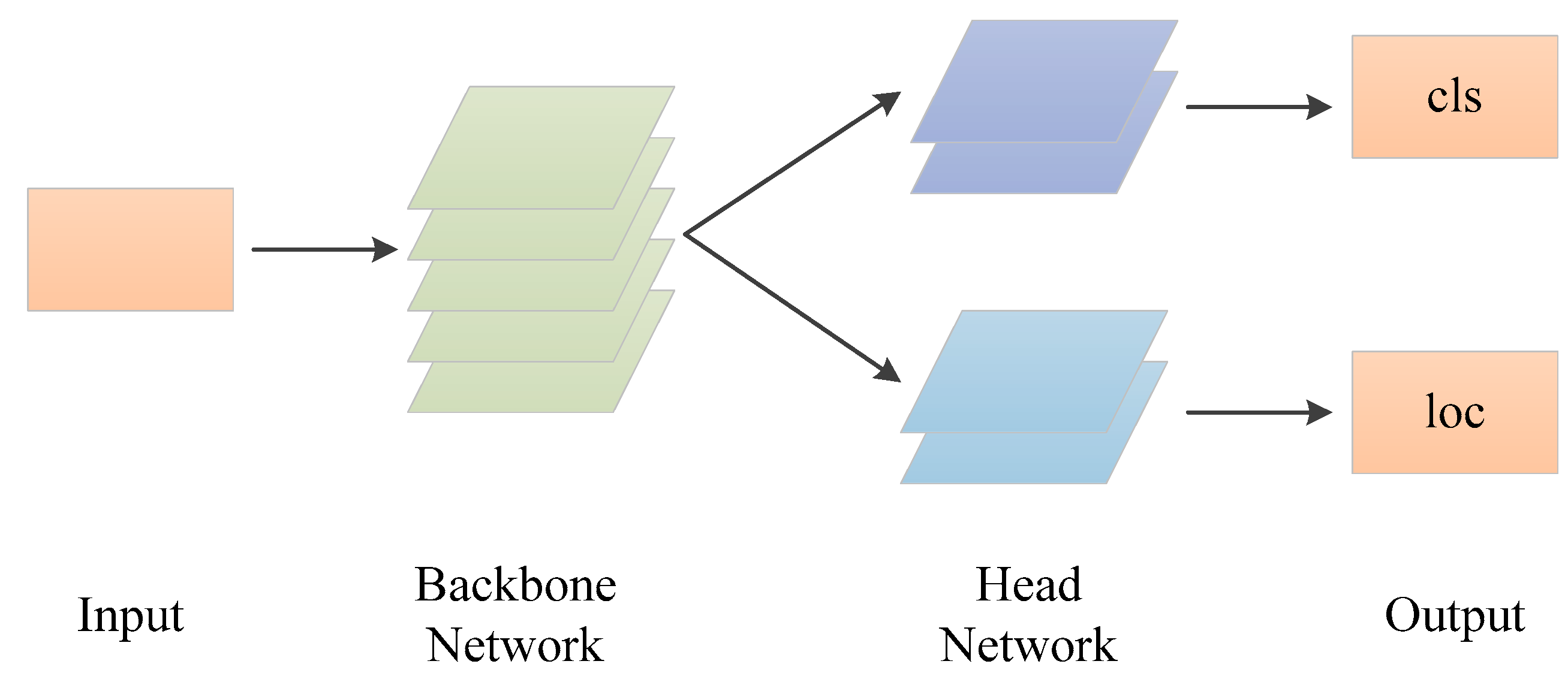


Figure 9 – YOLOv8 Architecture

Here, Would list the Yolov8 stages with brief details of each (CSPNet: A New Backbone That Can Enhance Learning Capability of CNN: This paper introduces Cross Stage Partial Network (CSPNet), a technique used in later YOLO versions for improved backbone performance.

Link: https://arxiv.org/abs/1911.11929

YOLOv8 Documentation: The official YOLOv8 documentation provides details on the architecture and components of the model, including the backbone.

Link: <https://docs.ultralytics.com/>)

The core component, in YOLOv8 is in charge of extracting details from the image, which are later utilized by following layers (, like the neck and head) to carry out object detection. The core processes the image across layers to generate feature maps that depict varying degrees of complexity.

The core of the system uses activation functions such, as Leaky ReLU or Mish to add linear elements to the model. This assists the network, in grasping patterns by moving beyond changes. These activation functions come into play following processes to identify which characteristics are triggered and propagated throughout the network.

The end result of the backbone comprises feature maps at scales, which are subsequently transferred to the neck and head of the network for additional refinement. These feature maps preserve details necessary, for precise object detection and categorization.

The second stage YOLOv8 is neck, its component crafted to optimize the scale feature maps produced by the backbone and refine them prior, to inputting them into the detection head. This step is vital in enhancing the model’s capability to identify objects of sizes and scales in an image. It guarantees that the model can grasp and merge both data and intricate details, which are crucial, for precise object detection. (FPN for Object Detection: This paper introduces the concept of Feature Pyramid Networks and its application in object detection.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

Link: https://arxiv.org/abs/1612.03144

Path Aggregation Network for Instance Segmentation: This paper details PANet, which is used for enhancing feature maps in the neck of object detection models.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

Link: <https://arxiv.org/abs/1803.01534>

Last stage is the YOLOv8 head, it plays a role, in converting the processed feature maps from the neck into detections. This involves forecasting the bounding boxes identifying the objects, in those boxes and determining confidence levels for each detection. It is the stage where both localization (identifying object locations) and classification (determining object types) tasks are carried out simultaneously.

The results, from the section show a collection of boxes that outline objects each labeled with a class and confidence level. These results usually follow the structure [center on x center, on y axis width, height, object score class scores]. Subsequently these estimations are employed to outline boxes around identified objects. Label them in the concluding image.

### 4.2 Dataset Construction

A new dataset tailored for PPE detection is created, comprising images captured for PPE. The dataset includes 1400 instances of PPE, featuring diverse types of PPE such as helmets, safety goggles, vests, gloves, and safety shoes. The data collection process, annotation, and preparation techniques are detailed, ensuring the dataset's comprehensiveness and relevance.

The dataset comprises images from multiple sources, including publicly available PPE datasets, images labeled with various PPE classes such as helmets, vests, gloves, safety shoes, and safety goggles.

Each image is annotated with bounding boxes around the PPE items, labeled according to their class, annotation and labeling for pictures using Roboflow website, where all instances annotated and labeled.





Figure 10 - Annotate

The dataset is split into three sections; Testing, Training and Validation, with each part constituting 10%, 70% and 20% of the dataset.

Below sample of training batch

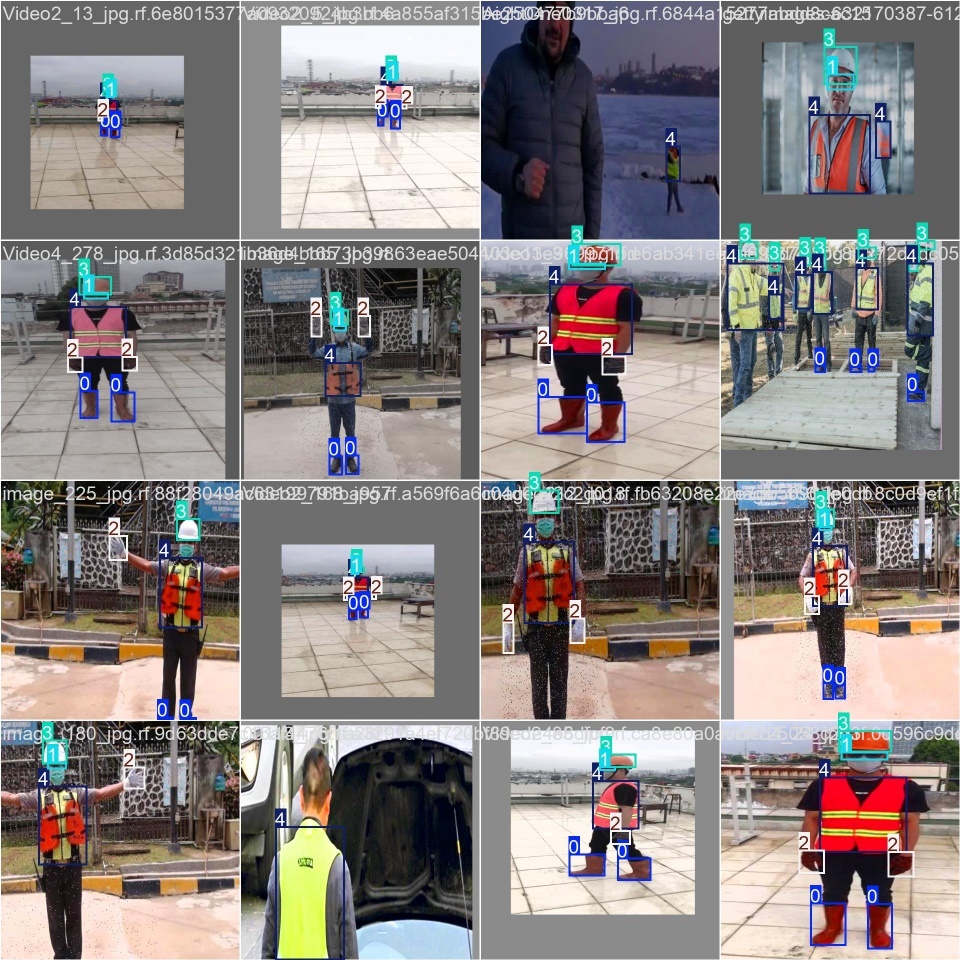


Figure 11 – training Images

Also, below picture for validation sample



Figure 12 – validation Images

Also, data sets contain instances count as below, show balancing in instances used to train the model

|  |  |
| --- | --- |
| Class | Images |
| all | 236 |
| boots | 212 |
| glasses | 176 |
| gloves | 205 |
| helmet | 211 |
| vest | 228 |

Table 4 - Instances count

Various augmentation techniques, such as rotation, flipping, scaling, and color adjustments, Noise Injection are applied to the dataset to increase its diversity and help the model generalize better to new data.

Noise Augmentation Sample as shown below in figure 13



Figure 13 – Noise Augmentation

Gray Scale Sample as shown below in figure 14



Figure 14 – Gray Scale Augmentation

Rotation & Flipping Sample as shown below in figure 15

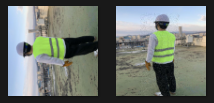


Figure 15 – Rotation & Flipping Augmentation

Scaling Sample as shown below in figure 16

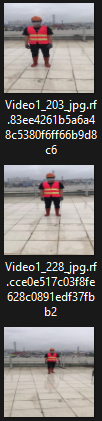


Figure 16 – Scaling Augmentation

### Training and Evaluation

YOLOv8 Model trained using Google colab platform using Graphical Processor Unit (GPU), These types of processors are widely used in computer vision tasks for object detection due to their ability to handle processing, which is essential for managing the complex calculations needed by deep learning models, below why GPUs are crucial in object detection.

1. Parallel Processing; Object detection entails recognizing and categorizing objects in an image requiring substantial computational power. GPUs can manage thousands of threads simultaneously making them well suited for the processing demands of deep learning algorithms this parallelism enables training and inference times compared to CPUs.

2. Throughput; GPUs provide high throughput, crucial for handling large datasets and high-resolution images commonly used in object detection tasks. This feature ensures that models can be trained rapidly and process images in time, which is vital for applications such as autonomous vehicles and real time surveillance systems.

3. Deep Learning Frameworks Optimization; Modern deep learning frameworks are tailored for GPU acceleration.

4. Memory Bandwidth; GPUs typically possess memory bandwidth, than CPUs enabling them to efficiently manage the extensive data volumes involved in deep learning tasks.

When it comes to object detection the ability of models to analyze images or video frames is essential for optimal performance.

5.Energy Efficiency; Despite GPUs requiring power compared to CPUs they prove to be more energy efficient when handling extensive computations. For example, training a learning model using a GPU is typically quicker and consumes less energy overall than using a CPU making GPUs better suited for energy demanding tasks such, as object detection.

References:

Jia, Y., Shelhamer, E., Donahue, J., et al. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. Proceedings of the ACM International Conference on Multimedia - This paper discusses the use of GPUs in training convolutional networks for image classification and object detection, showcasing the benefits of GPU acceleration.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems - This seminal paper on the AlexNet architecture highlights the critical role of GPUs in training deep networks, including their application in object detection.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - The YOLO (You Only Look Once) model is a real-time object detection system that heavily relies on GPU processing power, demonstrating the practical application of GPUs in object detection.

After collection of the PPE pictures and augmented, we started to learn the Model using YOLOv8, the models trained for 50 epochs (the number of times the network iterates over the data), we reach to 98.4% of mAP.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Images** | **Instances** | **Recall** | **mAP** |
| all | 236 | 1461 | 98.20% | 98.40% |
| boots | 212 | 434 | 100.00% | 99.50% |
| glasses | 176 | 179 | 93.70% | 95.90% |
| gloves | 205 | 384 | 97.10% | 97.70% |
| helmet | 211 | 225 | 100.00% | 99.50% |
| vest | 228 | 239 | 100.00% | 99.50% |

Table 5 – Showing instances and mAP & Recall metric.

### 4.4 Performance Metrics and Benchmarking

Evaluating the effectiveness of learning models in tasks like detecting Personal Protective Equipment (PPE) relies heavily on performance metrics and benchmarking. These metrics offer insights into how well a model can identify and locate objects in images enabling a thorough evaluation of its strengths and weaknesses.

Among the performance metrics is Mean Average Precision (mAP) which is widely used in object detection tasks. It measures the model’s precision across intersection, over union (IoU) thresholds and all classes providing a consolidated view of its performance. A higher mAP value signifies detection accuracy overall.

Calculating mAP involves determining the Average Precision (AP) for each class by integrating the precision recall curve and averaging these values across all classes. Another important metric is Precision, which indicates how accurately the model predicts outcomes.

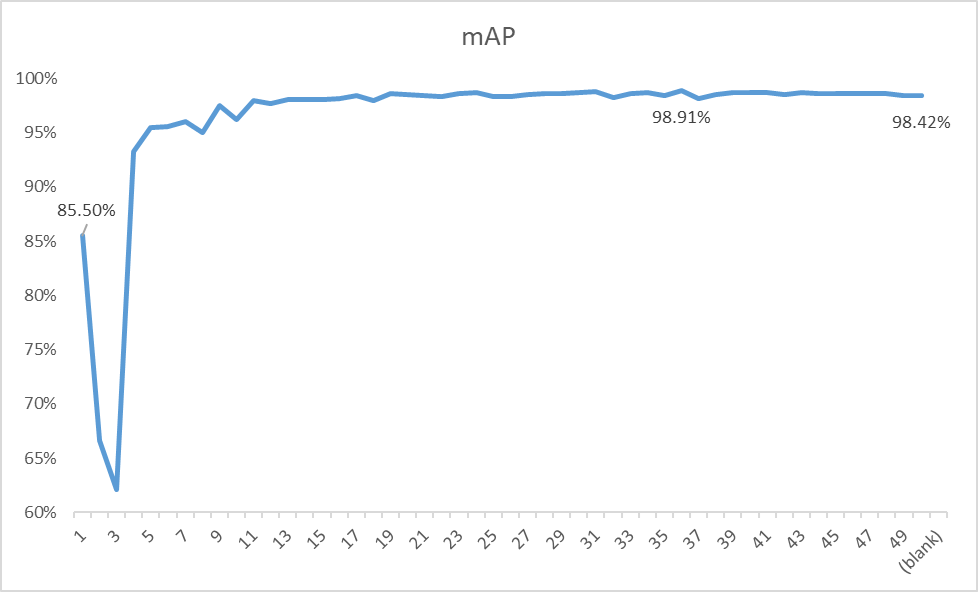


Figure 17 – mAP for YOLOv8 Object detection algorithm

The precision refers to the ratio of identified objects to the total number of predicted positives, including true positives and false positives. When precision is high it means that most of the detected objects are relevant, Precision= true positives (TP) out of all positive predictions. (TP+FP).

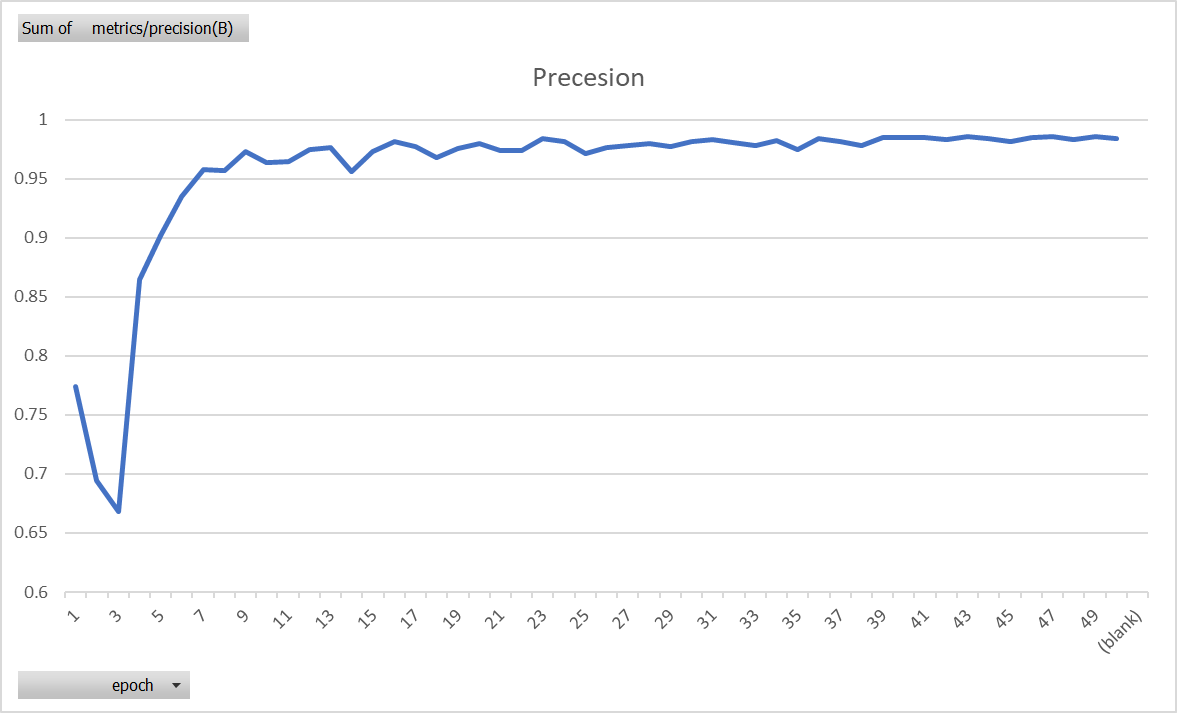


Figure 18 – precession for YOLOv8 Object detection algorithm

Another important metric is recall, which assesses the model’s ability to identify all objects within an image correctly. It is calculated as the ratio of positives to the total number of actual positives, including true positives and false negatives. High recall indicates that the model can detect most of the objects, Recall = true positives(TP) out of all predictions(TP+FN).

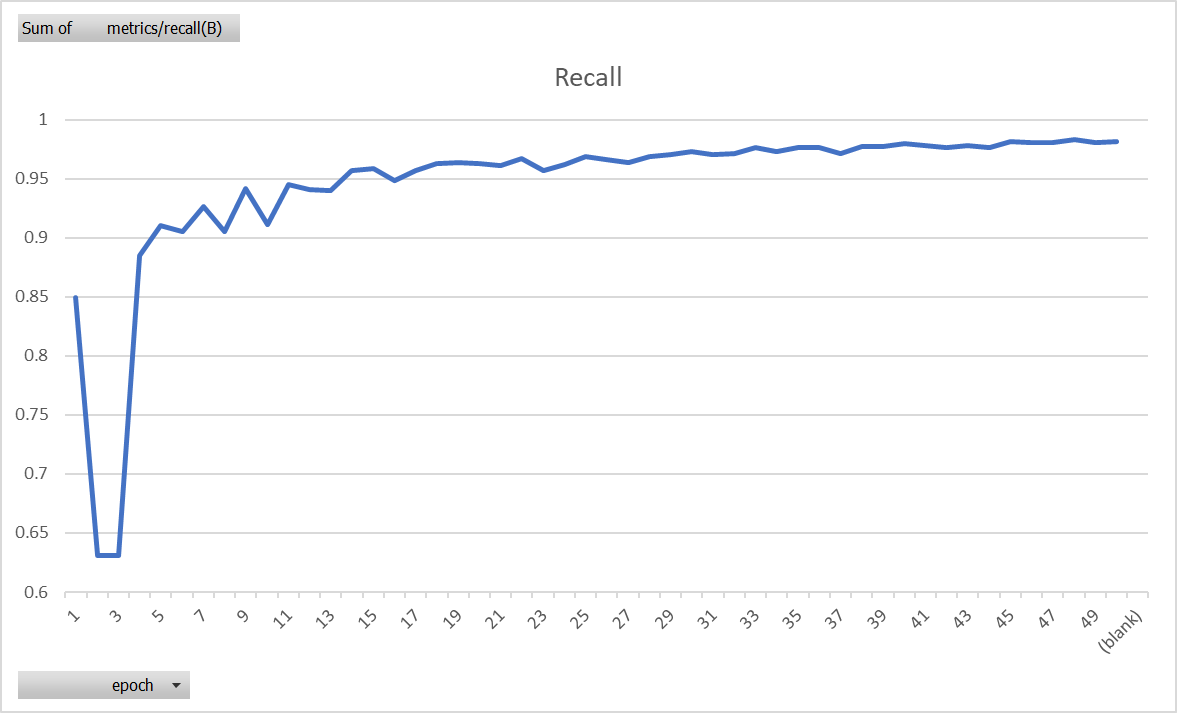


Figure 19 – Recall for YOLOv8 Object detection algorithm

The F1 Score combines precision. Recall into a single metric offering a balanced evaluation. This metric is especially beneficial when dealing with datasets that have imbalanced class distributions, in this training we achieve F1 score of 98.3%.

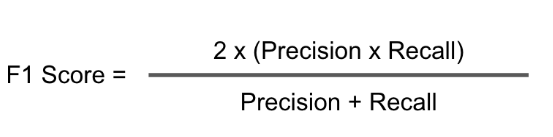


Figure 20 – F1 Score equation

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Images** | **Instances** | **F1 Score** |
| all | 236 | 1461 | 98.30% |
| boots | 212 | 434 | 99.70% |
| glasses | 176 | 179 | 94.84% |
| gloves | 205 | 384 | 97.55% |
| helmet | 211 | 225 | 99.95% |
| vest | 228 | 239 | 99.35% |

Table 6 – F1 Score Results

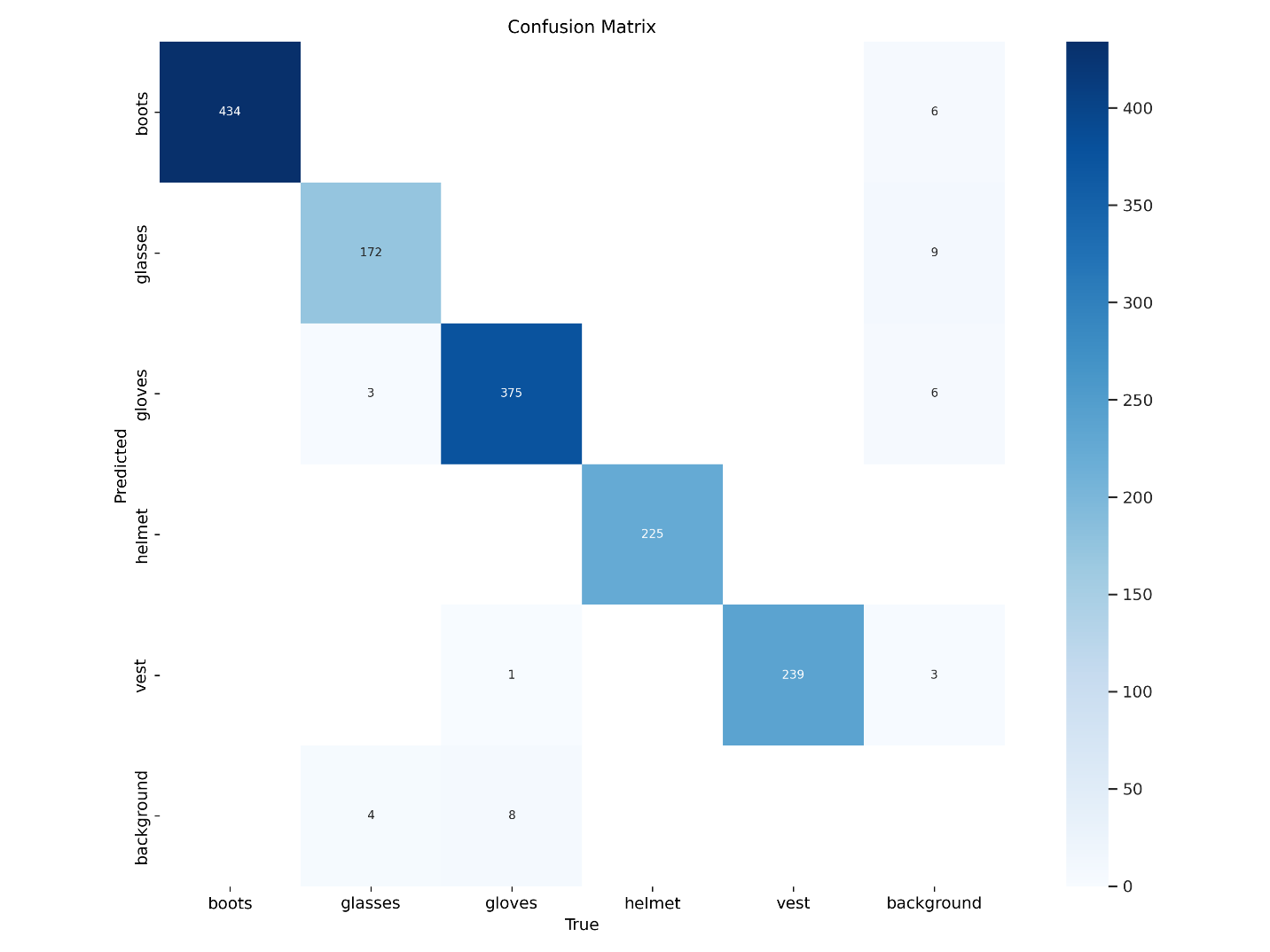
A confusion matrix is a table that showcases how well a classification model performs by comparing the actual and predicted classifications. It outlines positives, false positives, true negatives and false negatives for each category. This matrix is valuable in pinpointing where the model struggles like having many incorrect predictions for certain PPE items. It's great for pinpointing class performance issues.

Figure 20 – Confusion Matrix

### Summary

In essence it is crucial to utilize performance measurements and comprehensive comparisons when assessing the efficiency of a PPE detection system based on YOLOv8. Through an examination of these measurements the initiative can guarantee that the model not just fulfills but surpasses the criteria for timely, precise and effective PPE detection in industrial settings. This thorough assessment procedure ultimately plays a role, in creating an adaptable safety surveillance solution.

## Chapter 5 Experimental Results and Discussion

### Performance Evaluation

The YOLOv8 models performance, in detecting equipment (PPE) was evaluated using a mix of quantitative measurements and qualitative assessment. It achieved a Mean Average Precision (mAP) score of 98% showcasing its ability to accurately spot and categorize PPE items like helmets, gloves and vests in different industrial settings. The precision and recall metrics were also calculated, showing a precision of and a recall of 98.4% indicating its capacity to correctly identify PPE while minimizing errors. Furthermore the real time performance was gauged by measuring Frames Per Second (FPS) during video processing revealing a FPS of 1 affirming its suitability for time critical environments where swift detection is essential. Despite these findings some limitations were noted like misclassifications in messy scenes or challenging lighting conditions. In summary the YOLOv8 model shows performance positioning it as a tool for improving safety monitoring through automated PPE detection, in industrial settings.

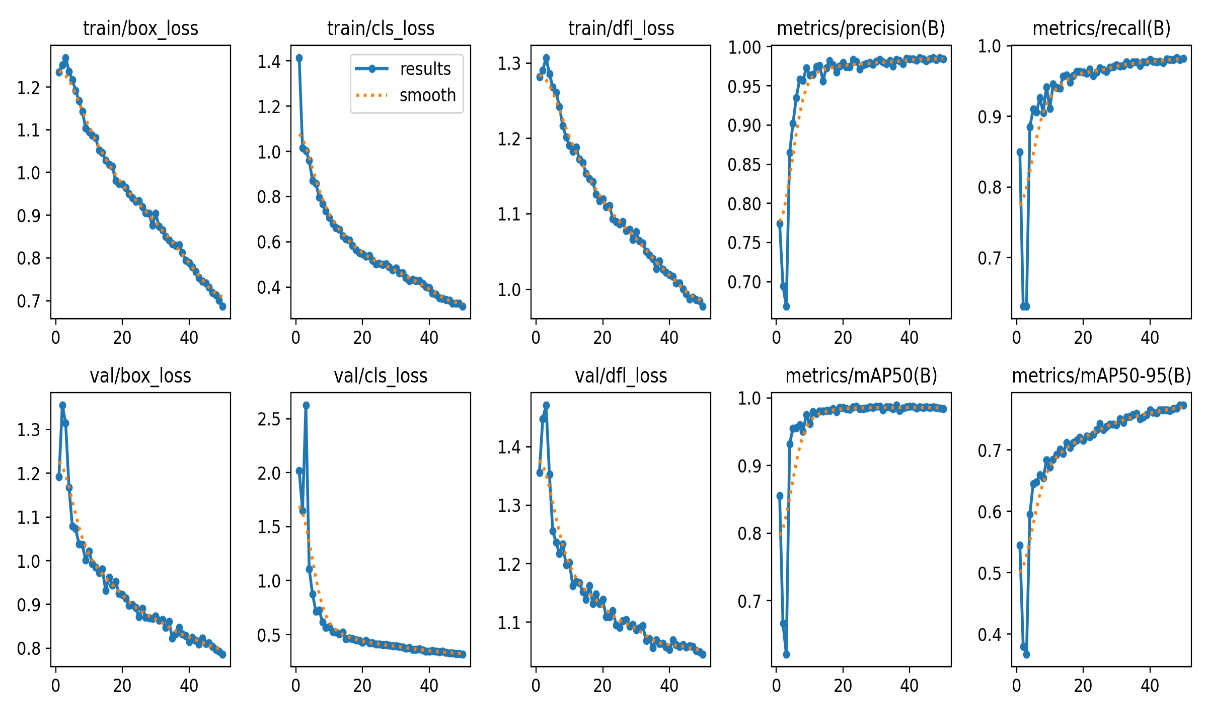


Figure 21 – All KPIs of the YOLOv8 Trained Model

### 5.2 Challenges and Limitations

While the YOLOv8 model showed performance in recognizing Personal Protective Equipment (PPE), One of the challenges faced during this study was the reliance on CPU processing, which caused delays in real time PPE detection. YOLOv8, similar to deep learning models is optimized for GPU acceleration, which significantly boosts processing speed through parallel capabilities. However, when using the model on a CPU the time taken for each frame increased considerably resulting in Frames Per Second (FPS) and slower overall performance. This slowdown affects the model’s ability to provide real time detection for safety monitoring applications that require immediate feedback to prevent accidents or ensure compliance. The limited computational capacity of the CPU also delays processing of high-resolution images or large data batches potentially causing delays and decreased detection accuracy due to limitations in using more advanced model configurations that require additional resources. Furthermore, longer processing times on a CPU can lead to increased latency reducing the effectiveness of the model in environments where quick responses are crucial for adapting to rapid changes. These hardware constraints highlight the necessity for robust computational resources, like GPUs or specialized AI accelerators to fully exploit YOLO capabilities and achieve real time performance as needed in practical scenarios.

## Chapter 6 Conclusion and Future Work

### 7.1 Conclusion

This research aimed to leverage Artificial Intelligence tools, in Computer Vision to enhance safety measures for humans and amplify the impact of AI in various industries. The paper explores into factors highlighting the significance of Personal Protective Equipment (PPE) in work settings affirming the need for the final model developed. The study proposes a solution to address issues highlighted in the introduction through real time identification of PPE noncompliance enabling measures to prevent incidents. The research holds implications for industries emphasizing sustainability and employee wellbeing alongside cost savings for companies. By employing the YOLOv8 architecture, the final model successfully detected five categories of PPE segments with an mAP50 accuracy rate of 98% on validation data. Furthermore, tuning model performance can be achieved by expanding training datasets to enhance detection accuracy

### 7.2 Future Work

In the future there are avenues to consider for enhancing the effectiveness and practicality of YOLO in detecting PPE. One potential area for improvement involves using a range of diverse datasets to train the model effectively across different industrial settings lighting conditions and object variations. This may include gathering data from various industries or supplementing existing datasets with synthetic images to replicate challenging situations. Another promising approach is optimizing the model for use on edge devices and systems with computational resources, In the future looking into how transfer learning or domain adaptation techniques can help the model adjust faster and better to situations without needing extensive retraining. This could make YOLO more adaptable across industries and scenarios. Handling challenges like CPU based deployment is crucial for enhancing YOLO effectiveness in detecting PPE in real time, within diverse and demanding environments.

## References

### [Abiodun, Oludare Isaac, et al., 2018]

Abiodun, Oludare Isaac, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. “State-of-the-art in artificial neural network applications: A survey”. Heliyon 4. 2018.

### [Abu Saa, Amjed, et al., 2019]

Abu Saa, Amjed, Mostafa Al-Emran, and Khaled Shaalan. “Factors Affecting Students’ Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques”. Tech Know Learn 24, 567–598., 2019.

## Appendices

Python Code

import cv2 # OpenCV library for image and video processing

from ultralytics import YOLO # Import YOLO model from ultralytics package

import cvzone # Import cvzone for drawing bounding boxes and text on images

import pandas as pd # Import pandas for data manipulation and storage

import time # Import time module for FPS calculation

# Load YOLO model from the specified file path

model = YOLO("./best.pt") # Replace with the path to your trained model

# Define class names corresponding to the classes in your dataset

classNames = ['boots', 'glasses', 'gloves', 'helmet', 'vest']

# Open the video file or capture from webcam

cap = cv2.VideoCapture("ppe.mp4") # Replace with the path to your video file or use 0 for webcam

# Create an empty DataFrame with specified columns to store detection results

df = pd.DataFrame(columns=['Class', 'Confidence', 'X1', 'Y1', 'X2', 'Y2'])

# Initialize variables to calculate Frames Per Second (FPS)

prev\_frame\_time = 0 # Time of the previous frame

new\_frame\_time = 0 # Time of the current frame

# Start a loop to process each frame in the video

while True:

new\_frame\_time = time.time() # Capture the current time for FPS calculation

# Capture a frame from the video

success, img = cap.read()

# Check if the frame was successfully captured

if not success:

print("Error reading frame from video stream.") # Print an error message if the frame is not captured

break # Exit the loop if there is an error

# Perform object detection on the current frame

results = model(img, stream=True) # Run the YOLO model on the image

# Iterate over each detection result

for r in results:

for box in r.boxes: # Iterate over each detected box

# Extract bounding box coordinates, confidence score, and class index

x1, y1, x2, y2 = int(box.xyxy[0][0]), int(box.xyxy[0][1]), int(box.xyxy[0][2]), int(box.xyxy[0][3])

conf = float(box.conf[0]) # Confidence score of the detection

cls = int(box.cls[0]) # Class index of the detected object

try:

# Check if the class index is within the valid range of class names

if 0 <= cls < len(classNames):

# Draw the bounding box and label on the image using cvzone

cvzone.cornerRect(img, (x1, y1, x2 - x1, y2 - y1))

cvzone.putTextRect(img, f'{classNames[cls]} {conf:.2f}', (max(0, x1), max(35, y1)), scale=1, thickness=1)

else:

# If the class index is out of range, display a warning and use a default label

print(f"Warning: Class index {cls} out of range. Using default label.")

cvzone.putTextRect(img, "Unknown Class", (max(0, x1), max(35, y1)), scale=1, thickness=1)

# Add the detection result to the DataFrame

df = pd.concat([df, pd.DataFrame({'Class': classNames[cls], 'Confidence': conf, 'X1': x1, 'Y1': y1, 'X2': x2, 'Y2': y2}, index=[0])], ignore\_index=True)

except IndexError:

# Handle any errors related to index out of range during detection processing

print("Error: Index out of range. Skipping detection.")

# Calculate FPS by taking the reciprocal of the time difference between frames

fps = 1 / (new\_frame\_time - prev\_frame\_time)

prev\_frame\_time = new\_frame\_time # Update previous frame time for the next iteration

# Display the FPS on the image

cv2.putText(img, f"FPS: {int(fps)}", (40, 50), cv2.FONT\_HERSHEY\_SIMPLEX, 1, (0, 0, 255), 3)

# Show the processed frame in a window

cv2.imshow("Image", img)

# Check if the 'q' key was pressed to exit the loop

if cv2.waitKey(1) & 0xFF == ord('q'):

break # Exit the loop if 'q' is pressed

# Write the collected detection results to an Excel file

df.to\_excel('detections2.xlsx', index=False)

# Release the video capture and close any OpenCV windows

cap.release()

cv2.destroyAllWindows()