

## Article

# Real-Time Personal Protective Equipment Non-Compliance Recognition on AI Edge Cameras

Pubudu Sanjeewani <sup>1,\*</sup>, Glenn Neuber <sup>2</sup>, John Fitzgerald <sup>1</sup>, Nadeesha Chandrasena <sup>1</sup>, Stijn Potums <sup>1</sup>, Azadeh Alavi <sup>3</sup> and Christopher Lane <sup>1</sup>

<sup>1</sup> Research and Development, Smart AI Connect, Brisbane 4215, Australia; john.fitzgerald@smartaiconnect.com (J.F.); nadeesha@smartaiconnect.com (N.C.); stan.potums@smartaiconnect.com (S.P.); chris.lane@smartaiconnect.com (C.L.)

<sup>2</sup> Research and Development, MaxusAI, Brisbane 4006, Australia; glenn.neuber@maxusai.com

<sup>3</sup> STEM College, School of Computing Technologies, RMIT University, Melbourne 3000, Australia; azadeh.alavi@rmit.edu.au

\* Correspondence: pubudu@smartaiconnect.com

**Abstract:** Despite advancements in technology, safety equipment, and training within the construction industry over recent decades, the prevalence of fatal and nonfatal injuries and accidents remains a significant concern among construction workers. Hard hats and safety vests are crucial safety gear known to mitigate severe head trauma and other injuries. However, adherence to safety protocols, including the use of such gear, is often inadequate, posing potential risks to workers. Moreover, current manual safety monitoring systems are laborious and time-consuming. To address these challenges and enhance workplace safety, there is a pressing need to automate safety monitoring processes economically, with reduced processing times. This research proposes a deep learning-based pipeline for real-time identification of non-compliance with wearing hard hats and safety vests, enabling safety officers to preempt hazards and mitigate risks at construction sites. We evaluate various neural networks for edge deployment and find that the Single Shot Multibox Detector (SSD) MobileNet V2 model excels in computational efficiency, making it particularly suitable for this application-oriented task. The experiments and comparative analyses demonstrate the pipeline's effectiveness in accurately identifying instances of non-compliance across different scenarios, underscoring its potential for improving safety outcomes.

**Keywords:** instance segmentation; object detection; edge computing; personal protective equipment; deep learning; real-time AI

**Citation:** Sanjeewani, P.; Neuber, G.; Fitzgerald, J.; Chandrasena, N.; Potums, S.; Alavi, A.; Lane, C. Real-Time Personal Protective Equipment Non-Compliance Recognition on AI Edge Cameras. *Electronics* **2024**, *13*, 2990. <https://doi.org/10.3390/electronics13152990>

Academic Editors: Sandra V.B. Jardim, Rolando Miragaia and José Carlos Bregieiro Ribeiro

Received: 21 June 2024

Revised: 15 July 2024

Accepted: 24 July 2024

Published: 29 July 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The construction industry is a constantly evolving sector with numerous advancements in equipment and technology. However, the construction industry still has a high number of work-related accidents and injuries. Nevertheless, given the number of avoidable accidents and fatalities that occur each year, paying attention to taking necessary actions to improve existing safety systems is important. The industry still relies heavily on labor and is subject to frequently changing working environments, including the involvement of numerous work groups and heavy machinery.

At least 60,000 fatal accidents occur on construction sites worldwide each year, according to estimates from the International Labor Organization (ILO) [1], highlighting the urgent need to improve safety at construction sites. As per the statistics provided, construction workers account for one fatal accident every ten minutes, and one in every six fatal accidents at work occurs on a construction site. A total of 417 construction employees died from workplace injuries during the years 2002–2003 and 2013–2014, according to data from the Construction Industry Profile [2]. This accounted for 14% of all workplace

fatalities in Australia during this time period. The overall death rate is equivalent to 2.24 fatalities for every 100,000 workers, which is 34% more than the national rate of 1.67. These shocking statistics highlight the need for implementing a solid safety management system to ensure the safety of construction workers.

The Australian construction industry has consistently experienced high rates of workplace accidents and fatalities, underscoring the urgent need for effective safety measures. A comprehensive understanding of the primary causes of accidents and the unique challenges faced by the industry is essential to mitigate these risks and enhance workplace safety for all Australian construction workers. Automated safety monitoring systems can significantly reduce the risk of accidents by continuously monitoring the work environment, detecting potential hazards, and ensuring compliance with safety regulations. Employers have both a legal and moral obligation to ensure the safety of their workforce. This includes providing appropriate safety equipment, such as personal protective gear, and fostering a safe working environment. Consequently, the development and implementation of real-time Personal Protective Equipment (PPE) non-compliance recognition on AI edge cameras has been primarily motivated by the need to ensure worker safety and reduce incidents on construction sites. AI edge technology is used in this context to bring real-time PPE non-compliance recognition directly to AI edge cameras. By deploying AI processing capabilities at the edge where data are generated, rather than relying solely on centralized servers, construction sites can achieve immediate detection and response to safety violations. This approach enhances worker safety by swiftly identifying instances where safety protocols, such as wearing proper PPE, are not being followed. This proactive monitoring not only helps in complying with safety regulations but also minimizes the risk of accidents, thereby safeguarding the well-being of construction workers.

The deaths and injuries caused in construction sites may occur due to body stressing, falls, trips, and slips, being hit by moving objects or falling objects, falls from a height, being struck by an object, electrocutions, vehicle collisions, and being caught in or in between an object. Accidents that involve head injuries are the most severe type of injuries that can happen on a construction site. Traumatic Brain Injury (TBI) is the most common type of head injury that occurs when the rapid acceleration or deceleration of the head causes the brain to move and collide with the skull. The most frequent causes of TBIs on construction sites have been identified to be falls and being hit by or against an object. In order to prevent TBIs, it is very important to wear a hard hat on construction sites, and require further enhancements in head protection gear, too. It is also necessary to wear high-contrast safety vests in order to locate people moving around the construction site. These vests can prevent injuries that may occur on the site due to the poor visibility of workers. Therefore, to mitigate the risk of deaths and injuries at work, it is recommended to use proper Personal Protective Equipment (PPE), such as hard hats and safety vests.

Australian regulations [3] mandate that safety hard hats be worn on sites when there is a risk of being hit in the head, whether by falling objects or from a collision with a fixed object, recognizing head injuries as a significant factor influencing the safety of the construction site. The Australian Work Health and Safety Strategy 2012–2022 [4] made it a priority to address the above issues and identified the construction industry as a national priority for prevention activities in Australia. According to Australia standard AS/NZS 1800:1998 [5], which specifies all the requirements for protective helmets in the workplace, different types of protective helmets are available in the industry. Similarly, Australian Standard High Visibility Safety Garments AS/NZS 4602:1999 [6] covers three types of garments: daytime wear, night-time wear, and those that can be worn day and night. The jackets are designed with a safety reflective strip on the jacket.

Construction safety management has traditionally been extremely hard due to the complex motion and interaction of people, goods, and power. Traditionally, a foreman and additional safety officers were responsible for directly supervising workers and enforcing safety regulations on construction sites. However, due to the sheer disproportion

in numbers, continual supervision on construction sites is not practicable. Because there are too many workers to monitor in the field, labor management is tedious and time-consuming. Accidents may occur as a result of inadequate safety instruction and supervision, unstable loads being lifted, or unsafe structures. The site monitoring task performed by site safety officers traditionally can be improved by using innovative methods to monitor the workers more comprehensively.

Several techniques based on wearable sensors and vision monitoring have been utilized to ensure the proper use of PPE. There are numerous methods in the industry utilizing electronic equipment like Radio Frequency Identification (RFID) tags [7] and sensor-integrated PPE equipment with Bluetooth transmitters [8] to decrease the expenses and time consumed for inspecting the usage of PPE. Hardware-based approaches are not suitable for real-time monitoring because of their high installation and maintenance costs, low accuracy, and slow speed making them unsuitable for real-time monitoring. These hardware sensors and receptors are more expensive to maintain and are subject to damage and wear and tear.

Using computer vision systems is another straightforward way to operate continuously and eliminate any human issues like fatigue, inattention, and illness in controlling PPE wearing. However, it is not sufficient to only identify workers, hard hats, and vests. It is necessary to determine the relationship between these instances, too. In actuality, the main issue is not locating those who appropriately wear PPE but rather locating those who disregard safety regulations by not utilizing it. Computer vision techniques have demonstrated their efficacy in rapidly and conveniently extracting pertinent data from construction sites. These techniques encompass the identification and monitoring of workers and equipment, along with numerous other Information Technology (IT) and computer-based tools extensively employed within the construction industry to automate diverse processes. However, this area of study is still in its infancy.

As a result of disparities in orientation, color, background contrast, image resolution, and on-site lighting conditions, current systems either lack comprehensive testing under escalating levels of difficulty or are victims of over-prediction and erroneous identification of undesired objects such as hard hats or vests. Additionally, none of the earlier investigations have addressed the time efficiency of the detection method to be used in real-time applications. After all, especially in safety applications, the computational speed of the algorithm is just as crucial as its accuracy. Cutting-edge surveillance systems with more advanced imaging and post-processing technologies are constantly in demand. Digital video surveillance systems entirely rely on human operators to identify threats. They are only intended to offer the infrastructure needed to record, store, and distribute video. Currently, government departments and private entities manage very large camera IP networks for security and public safety, often without the management tools or consideration given to AI implementation. Monitoring surveillance footage manually requires a lot of labor, and manual video searches take too long to yield the crucial data needed to support investigations. Therefore, there is a greater demand for automation in this area.

This research seeks to automate a part of the construction safety inspections by automatically identifying persons not wearing hard hats and safety vests using real-time video footage in order to alleviate the aforementioned limitations using recent advancements in computer vision. Elhanashi et al. [9] discuss the integration of deep learning into IoT, highlighting techniques and challenges for real-world applications. Their insights emphasize the relevance of leveraging AI and IoT technologies, including edge computing, to address complex safety challenges in industries like construction. In this context, a smart surveillance solution that detects workers without hard hats and vests in construction sites has been developed. The proposed system receives data from iPRO (formerly Panasonic) AI-enabled edge-capable IP security cameras at production or construction sites and rapidly identifies the absence of a protective hard hat on a person's head and a safety vest worn in real-time while scanning the site. The proposed system does not rely on the implementation of a Video Management System (VMS) and can run alongside or in conjunction.

Then, the system sends alerts or instant notifications to the management teams of the work rules violation so that they can take necessary action to prevent any injury. The proposed system also offers compliance monitoring with safety rules in places where equipment is in use, installation work is performed or in places of unloading/loading of raw materials takes place, and control of discipline at the work site.

The original contributions of this paper are as follows:

1. Innovative end-to-end PPE detection model: In contrast to existing expensive sensor-based and multistage vision-based approaches, this paper introduces a cost-effective, single-stage end-to-end model for detecting PPE usage. This model demonstrates robustness and effectiveness in challenging real-world scenarios, providing a streamlined solution that simplifies deployment and maintenance.
2. Enhanced dataset augmentation techniques: The paper employs advanced augmentation techniques to expand the dataset size and improve detection accuracy for small objects. These techniques also enrich the model with a greater amount of semantic information from features, significantly enhancing its ability to recognize and classify PPE in diverse environments.
3. Automatic instance segmentation and classification: The proposed method utilizes an SSD MobileNet V2-based technique for automatic instance segmentation and classification of PPE from real-time construction surveillance video data. This approach ensures precise detection and classification, contributing to improved safety monitoring on construction sites.
4. Edge deployment for real-time processing: The model is deployed on the edge, where CCTV cameras act as sensors, processing video feeds in real-time without the need for connection to a Video Management System (VMS). This deployment reduces dependency on traditional cloud services, which often require high-speed internet and generate vast amounts of data for analysis, thus providing a more efficient and scalable solution.
5. AI-enabled edge processing: By enabling AI processing directly on CCTV cameras, the need for local or cloud-based processing is eliminated. The approach not only reduces latency but also enhances the system's reliability and responsiveness, making it suitable for real-time PPE detection and alerting. The model has been deployed on the edge where the CCTV camera acts as a sensor that does not need to be connected to a VMS with efficient computation and streaming of multiple high-quality video feeds in real-time, replacing the traditional cloud services, which require lightning-fast internet connection while providing an awful amount of data to analyze.
6. Integration with external sensors: The AI-enabled CCTV cameras are equipped with network and digital I/O outputs, allowing for the connection of external sensors such as sirens or warning lights. This integration facilitates immediate on-site alerts and enhances the overall safety infrastructure.
7. Extensive experimental validation: The proposed model undergoes extensive experimentation and comparative analysis using real-world construction video data. This rigorous testing demonstrates the model's effectiveness and robustness, providing solid evidence of its practical applicability and reliability in diverse construction site conditions.
8. Comparative analysis of real-world data: The paper presents a detailed comparative analysis of experimental results obtained from practical surveillance data on an on-site construction project. These results validate the proposed network's effectiveness and robustness, showcasing its superior performance in real-world scenarios compared to existing methods.

The proposed smart surveillance solution addresses several key limitations of current safety monitoring systems in construction environments. By automating safety inspections through real-time video analysis and computer vision advancements, the system reduces reliance on manual monitoring prone to human error, inconsistency, and fatigue.

Interpreting data from traditional systems can be complex and time-consuming, requiring skilled personnel for analysis and decision-making. Integration with AI-enabled edge-capable cameras allows for decentralized processing, eliminating the latency associated with centralized systems and enhancing responsiveness to safety violations. By leveraging AI-edge technology, the system processes data directly on the CCTV cameras, eliminating the need for extensive cloud infrastructure. The model has been deployed on the edge where the CCTV camera acts as a sensor that does not need to be connected to a VMS with efficient computation and streaming of multiple high-quality video feeds in real-time, replacing the traditional cloud services, which require a lightning-fast internet connection while providing an awful amount of data to analyze. This approach not only improves coverage and accuracy in detecting non-compliance with safety regulations, enhancing reliability and making it suitable for real-time PPE detection, but also mitigates the high costs and complexity associated with traditional sensor-based solutions which are expensive to install, maintain, and upgrade. The cost factor can limit widespread adoption, particularly in smaller firms or projects with limited budgets. By providing immediate alerts of safety equipment non-compliance and enabling proactive management responses, the system facilitates timely interventions to prevent accidents, thereby significantly enhancing overall safety management in construction sites. Furthermore, the system enhances compliance monitoring in areas involving equipment use, installation work, and material handling, thereby bolstering discipline and safety standards at construction sites.

In the following sections, we compare and contrast different techniques and optimizations for the solution on edge devices aiming for commercial deployment. The rest of this paper is organized as follows. In Section 2, a literature review regarding deep learning in the context of construction safety and hard hat and safety vest non-compliance detection is presented. Section 3 describes the proposed solution, its implementation, training, and testing procedures. Experiments and results are presented in Section 4, while in Section 5, result analysis and discussion are presented. In Section 6, comparisons with other methods have been made. Finally, conclusions drawn from that with suggestions for future work are presented in Section 7.

## 2. Related Works

This section discusses the related works and contributions to video surveillance applications in identifying workers not wearing hard hats and safety vests. Software solutions have utilized computer vision techniques to identify instances of non-compliance with PPE in two distinct methodologies: handcrafted feature extraction systems and deep learning approaches. Handcrafted feature extraction systems exhibited commendable performance levels but encountered limitations in accurately detecting PPE negligence across various positions and lighting conditions. On the other hand, deep learning approaches demonstrated superior accuracy in identifying PPE usage and negligence, albeit with slower performance, rendering them unsuitable for real-time monitoring applications. Traditional computer vision methods mostly relied on handcrafted features, which could be processed with machine learning algorithms. Several traditional computer vision methods such as Speed-Up Robust Features (SURF) descriptor, template matching technique using Histogram of Oriented Gradients (HOG) as features in cascade classifier [10], background subtraction [11], ViBe background modeling [12], and contour cues [13] have been used in PPE detection. However, these methods were vulnerable to weather and lighting conditions changes.

Several deep learning-based techniques have been used to detect PPEs, as shown below. Wójcik et al. [14] proposed a hard hat-wearing detection system based on head key point localization and achieved Microsoft Common Objects in Context (MS COCO) style overall Average Precision (AP) of 67.5% with class-specific AP for hard hat non-wearers of 64.1%. However, the proposed system did not work in real-time and did not have the ability to identify workers who broke the rules so they could be reprimanded, fined, or sent to additional OHS training. Shrestha et al. [15] proposed a hard hat detection system

for construction safety visualization, which failed when the CCTV camera could not capture front-facing images when the worker faces down while working and when a worker moves quickly. When the contrast between the hard hats and the background color was not adequately high, the edge detection program could not give a clear outline of the hard hats. Daghan et al. [16] proposed a deep learning model using face recognition and object detection for detecting PPE and obtained a Mean Average Precision (mAP) of 0.95. Mneymneh et al. [10] evaluated existing computer vision techniques in rapidly detecting hard hats on indoor construction sites and highlighted the potential of cascade classifiers. Kamal et al. [17] deployed an SSD MobileNet model on edge computing devices, Nvidia Jetson TX2 board, and nano board in real time for construction safety surveillance and received better performance without a considerable drop in accuracy, with fewer Frames per Second (FPS). Yu et al. [18] proposed a protective equipment detection algorithm fused with apparel check in electricity construction with performance enhancements in terms of precision by 2.4% than the state-of-the-art technique. Zaabi et al. [19] proposed an automatic site inspection system in construction sites using deep learning SqueezeNet pre-trained network to detect two main violations: health and safety and electrical hazards, and achieved accuracies of 91.67% and 92.86%, respectively. Zhang et al. [20] proposed a construction worker hard hat-wearing detection system based on an improved BiFPN and achieved an mAP of 87.04%, which outperformed several state-of-the-art methods on a public dataset. Nath et al. [21] proposed a deep Convolutional Neural Network (CNN)-based Generative Adversarial Network (GAN) to enhance image quality for fast object detection in construction sites. The findings indicated that the AP of pre-trained models employed to detect various objects, including buildings, equipment, workers, hard hats, and safety vests, can be enhanced by as much as 32% without compromising the overall real-time object detection processing time. Liu et al. [22] detected workers wearing safety helmets based on YOLOv4 in real-time and achieved an accuracy of 93% on 9986 hard hat datasets. Even though the detection accuracy was high, and the detection speed was fast, individuals positioned at a significant distance from the fixed camera missed the inspection in actual applications. Filatov et al. [23] developed a hard hat-wearing monitoring system using deep neural networks and surveillance cameras with high inference speed and achieved an F1-score of 0.75. The combined system comprising the SqueezeDet and MobileNets neural networks exhibited a notable 9% increase in the precision of safety adherence detection compared to a single SqueezeDet system. Pardhi et al. [24] developed and tested a hard hat acquisition tool using image processing techniques. However, the system failed when cameras could not take forward-looking images and when an employee was moving too fast. When the difference between the hard hats and the back color was not high enough, the edge access system could not provide a clear framework for hard hats. Kawade et al. [25] proposed a real-time construction safety equipment detection system based on YOLOv3 with an excessive precision of 98.32% and a recall rate of 92.15%. Tyagi et al. [26] proposed a multiple safety equipment detection at active construction sites using effective deep learning techniques and achieved a precision of 0.57, a recall of 0.72, an mAP of 0.63 (at 0.5), and an f1-score of 0.621. Zhou et al. [27] proposed safety helmet-wearing detection and recognition based on YOLOv4 and improved the detection accuracy of the YOLOv4 network architecture on the VOC2028 dataset with comparable training speed. The proposed system was more sensitive to smaller objects while reducing the impact of class imbalance by intensifying the network's focus on rare classes and adding some additional parameters to solve the imbalance between classes.

Most of the above applications were limited to a smaller number of PPE and had performance issues when deployed in real-world environments in real-time. Therefore, in order to mitigate risks within construction sites, there exists an urgent need for the development of a system that effectively surmounts the aforementioned drawbacks while providing remarkable speed and accuracy.

Sandler et al. [28] introduced MobileNetV2, a novel mobile architecture that significantly enhanced performance across various tasks and benchmarks while accommodating diverse model sizes. It achieved efficient memory usage during inference by leveraging standard operations compatible with all neural frameworks. For ImageNet, MobileNetV2 set new benchmarks across multiple performance metrics. In object detection, particularly on the COCO dataset, the architecture, when paired with the SSDLite detection module, surpassed state-of-the-art real-time detectors in both accuracy and model complexity. SSD MobileNet V2 was used in various applications, including sign language detection [29], face mask detection [30], real-time cat detection [31], river trash sorting [32], and road object detection [33], with speed and accuracy improvements over state-of-the-art techniques. Due to these benefits, SSD MobileNet V2 was chosen as the backbone of the proposed methodology.

### 3. Proposed Methodology

The proposed solution makes use of CCTV cameras to process the video directly on the camera at the construction sites to provide a safety surveillance system that lowers the cost and labor required to check on workers' use of PPE. Construction personnel who are observed without the required safety gear, specifically hard hats and safety vests, are detected by AI-enabled edge-capable IP security cameras locally on the construction sites. The proposed system detects one or more people within an image or a video frame and returns a box around each person, along with the location of the people and confidence scores. The model's goal is exclusively to recognize the existence and location of people in an image and classify people wearing hard hats and vests. The model does not attempt to discover identities or demographics and does not attempt to discover any other characteristics of the people in the video. A monitoring dashboard that has camera status and camera-related details, model operational status, inference stats, comprehensive compliance logs of the deployed model, and the latest data output, including runtime history, is provided for the end user. The proposed system was trained on four main classes: people with separate classes for hard hat only, vest only, hard hat and vest, and no safety gear. Each class was assigned a separate RGB color value based on the traffic light color schema (from green: OK, yellow/orange: attention, red: danger or alert) to visually aid during the data labeling process. The class descriptions and the colors used for each class are mentioned in Table 1. For each person detected in a photo or video, the model outputs bounding box coordinates, hard hat only, vest only, hard hat and a vest, or no safety gear class, and detection and landmarking confidence scores. The proposed methodology is illustrated in Figure 1 below.

Helper classes were used during labeling and AI model creation to identify data bias and mitigate class imbalances to improve evaluation and out-of-distribution test performance. Helper classes used in the system are presented in Table 2 below, along with the class descriptions.

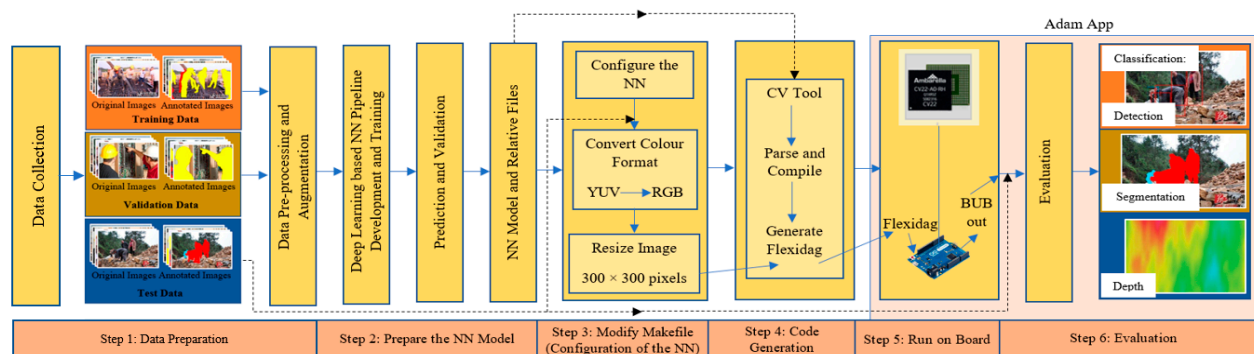


Figure 1. Proposed methodology.

**Table 1.** AI Detection Classes—Main.

ID	Display Name	Description	Color
1	Person: Hard hat and vest	Person wearing both a hard hat and high visibility safety vest with reflective stripes	Green
2	Person: Hard hat only	Person wearing only a hard hat (without high visible safety vest with reflective stripes)	Yellow
3	Person: Vest only	Person wearing only high visible safety vest with reflective stripes (without a hard hat)	Orange
4	Person: No safety gear	Person wearing no hard hat and no high visibility safety vest with reflective stripes	Red

### 3.1. Dataset Preparation

The model was trained over a subset of 1859 images (~70% of the original dataset) and validated on a validation dataset of 794 images (~30% of the total images). The training dataset contained samples with different images and human characteristics, including derived characteristics (human size, human posture, orientation, blur, lightness, and occlusion) and human demographics (human-perceived gender presentation, age, hair, and skin tone).

**Table 2.** AI Detection Classes—Helper.

ID	Display Name	Description	Mapping Main ID
5	Person: Vest and other head-gear	Person wearing only high visible safety vest with reflective stripes and headgear other than a hard hat, e.g., cap, beanie, hoodie, hat, etc.	3
6	Person: Other headgear only	Person wearing headgear other than a hard hat, e.g., cap, beanie, hoodie, hat, etc., and wearing no hard hat and no high visibility safety vest with reflective stripes	4
7	Not used: Hard hat	Hard hat instance, can be worn by a person or lying around	—
8	Not used: Vest	Vest instance can be worn by a person or lying around	—
9	Not used: Safety goggles	Safety goggles instance can be worn by a person or lying around	—
10	Not used: Safety boots	Safety boots instance can be worn by a person or lying around	—

#### 3.1.1. Data Collection

The training and evaluation datasets contained hard hat and vest images accumulated from several assets, such as:

- Images extracted from videos captured by a regular iPRO surveillance camera placed indoors. The surveillance camera was placed in an office environment as well as in a private apartment in Australia.
- Images extracted from videos captured by mobile phone cameras (e.g., iPhone 13 Pro Max, Samsung A52, Samsung Galaxy A52) while walking indoors and outdoors in Queensland, Australia.
- Images in the Hard Hat Workers Dataset [34] which comprises an object detection dataset specifically curated for workplace environments where the use of hard hats is mandatory. The dataset annotations encompass instances of workers wearing hard hats, as well as instances of individuals without hard hats, including annotations for



‘person’ and ‘head’ when an individual may be present without the required head protection.

- Images captured by digital cameras from actual construction sites.
- Images downloaded from the internet (natural photographs shared on Flickr, etc.).
- Images extracted by videos on the internet.

The training and evaluation data contained photographs captured from various settings, both indoors as well as outdoors. The indoor photographs included office environments, private apartments, shopping malls, etc. Outdoors included construction sites, office complexes, roads, shopping malls, etc. Both landscape and portrait images were incorporated into the dataset. These images were generally acquired from eye level. Image processing techniques, such as background blurring for portrait mode and filters, were not applied when capturing images. Data were stored on AWS S3 with strict access management to ensure that personal data were secure and used only for specific purposes. Some of the data contained staff from Smart AI Connect and KJR, Australia, who acted as PPE-wearing compliant and non-compliant construction workers and they have provided consent for their data to be used by the model.

As the trajectory of workers is random and the surveillance cameras are tuned for distinct resolutions, different distance and height situations were taken into consideration in experiments to validate the robustness of the proposed pipeline. A camera was mounted at different heights, and the optimal height to mount the camera to obtain the CCTV feed was observed. The model did not recognize human objects at 2.2 m in height. Between 1.5 m and 1.75 m in height, human objects were identified when the horizontal distance between the camera and the human object was more than 2 m in a visible image. The average height of a human being usually lies between 1.5 m and 1.75 m. Therefore, the camera was mounted at eye level with default settings, with the optimal distance of the subject from the lens being between 2 m and 10 m. Testing was performed in a large space where subjects could approach the camera from > 30 m away. Live inference detection and confidence were reviewed as the subject approached the camera, and the ranges chosen above reflected the best detection and highest accuracy. The subject then walked away from the camera to confirm the measurements. A number of different PPE combinations were used to further verify the optimum distance and height of the camera. Double checking was performed by having the subject walk backward and forward across the field of view of the camera at increasing 2 m distances. An optimal distance of 2 m–10 m was selected to mount the camera by comparing different accuracies obtained by different camera locations.

Default camera settings were used in the camera after testing several camera settings to see how they affected the model inference. The test procedure was as follows. A person stood still in front of the camera, and another person observed the live class detection and confidence values while recording the screen. The second person changed camera settings such as super dynamic on/off, intelligent auto on/off, stream settings such as frame rate, image quality 1 fine/0 super fine, smart VIQS on/off, and smart picture control on/off between each test and observed whether the class or confidence changes significantly. However, none of the settings made any noticeable difference to inference performance. Therefore, the default camera settings were used in the research. Figure 2 below shows some of the image samples for each object in the dataset.



**Figure 2.** Sample object samples cropped from the original images in the dataset for different classes.

### 3.1.2. Data Annotation

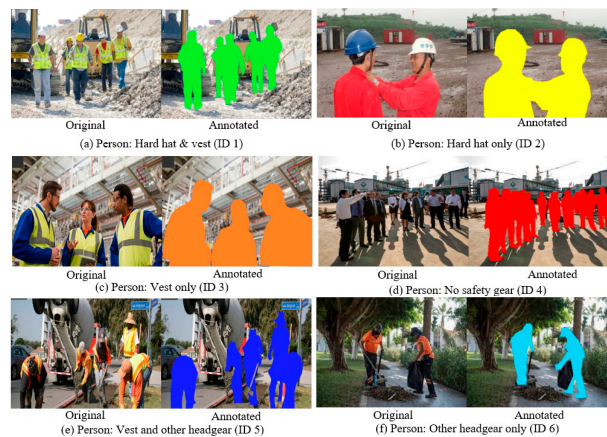
Training, validation, and test images were annotated on the MaxusAI™ LabelEngine platform premium version (MaxusAI: Brisbane, Australia), an end-to-end, in-browser, no-code, and AI-assisted active learning toolset developed by MaxusAI [35], Brisbane, Australia. LabelEngine has been developed to label and fully segment images and videos to create high-fidelity datasets for computer vision. The LabelEngine has the ability to label local images and videos directly in the browser without the need to install additional software or have development skills. Images were labeled only once with the help of objects/classes, hierarchies, and projects. Data were labeled more efficiently with advanced built-in features available in the LabelEngine environment, such as AI auto-segmentation, active learning, AI-enabled search, AI uncertainty ranking, and label analytics to monitor the distribution of annotations to manage and avoid class imbalances and prioritize the manual effort to maximize information gain to the neural network. All annotations were, in the first step, reviewed, and in the second step, validated by a trusted labeling team to ensure high quality and correctness with the associated class label in the AI training and test pool. Images lacking the proper presence of the objects under consideration were excluded from the dataset, as the resolution of this data complication was anticipated through the implementation of data augmentation techniques later. Figure 3 below shows some sample annotations representing each main class.

During data preparation, a more complex separate mask instance segmentation model learns in the background to assist the annotator with detailed predictions that are converted into editable, more detailed polygons.

### 3.1.3. Pre-Processing

Several pre-processing techniques were applied to the data before input to the system. The input images were resized to  $300 \times 300$  pixels before being fed into the model. This standardizes the image dimensions, ensuring consistency and compatibility with the input requirements of the SSD MobileNet V2 model. Resizing also helps reduce computational complexity and memory usage during model training and inference. The images captured were subjected to augmentation techniques to ensure the inclusion of diverse variations that are expected to occur in real-world scenarios. Augmentation techniques such as SSD random crop, random horizontal flip, padding, jpeg compression and darkening and brightening/color modification, blur, as well as randomized shift, scale, and rotations (between  $-15^\circ$  and  $15^\circ$ ) were applied. However, filtering, such as dropping images without faces, was not applied. SSD random crop involves randomly selecting a portion of the image and resizing it to the target dimensions ( $300 \times 300$  pixels). This technique helps make the model robust to variations in object positions and scales within the image. The random horizontal flip technique flips the image horizontally with a certain probability and introduces variability in the dataset, allowing the model to learn features invariant

to left–right orientation. Padding adds borders to the image, which can be useful in ensuring that objects near the edges are included in the input images after transformations. This helps prevent information loss due to cropping. Applying JPEG compression simulates the effects of image compression, which is common in real-world scenarios where images are often compressed to save storage space. This technique helps the model learn to handle various compression artifacts. Darkening and brightening techniques involve adjusting the brightness levels of the image to simulate different lighting conditions. This ensures that the model can perform well under varying illumination. Color modification involves changing the color properties of the image, such as hue, saturation, and contrast. This helps make the model invariant to color changes that might occur due to different lighting environments. Applying blur to the images simulates the effects of camera shake or focus issues. This helps the model learn to detect objects even when the images are not perfectly sharp. Randomly shifting the image horizontally or vertically by a certain number of pixels helps the model learn to detect objects regardless of their position in the image. Random scaling involves resizing the image by a random factor. This helps the model learn to detect objects at different scales. Rotating the image by a random angle between  $-15^\circ$  and  $15^\circ$  helps make the model robust to slight rotations that might occur in real-world images. Images were not filtered to exclude those without faces. This decision was likely made to ensure the model is exposed to a diverse range of scenarios, including those without any detectable faces. This helps train a more generalizable model that can handle various real-world conditions. These pre-processing techniques collectively enhance the robustness and generalizability of the model by exposing it to a wide range of variations that are expected in real-world scenarios. By augmenting the dataset with diverse transformations, the model becomes better equipped to handle different object positions, scales, lighting conditions, and image qualities during inference.



**Figure 3.** Original and annotation samples for each class.

### 3.2. Training

A handful of AI CV models that iPRO/Ambarella CV Tool can convert were identified for determining the most appropriate model for the application. Among them, YOLO v5 and SSD MobileNet V2 are deep learning-based object detection algorithms with high accuracy and speed that allow obtaining detection in real-time speed using iPRO cameras. The system was tested with advanced object detection models such as YOLO v5 and SSD MobileNet V2, and it identified that several models could be used to ascertain the utilization or disregard of PPE. The selection of YOLO v5 and SSD MobileNet V2 models for comparison was driven by their reputations for high accuracy and speed, making them suitable candidates for real-time object detection applications. Both models have demonstrated exceptional performance in previous studies, with YOLO v5 known for its accuracy and SSD MobileNet V2 recognized for its efficiency on resource-constrained devices.

[29]. Given the necessity for real-time PPE detection using iPRO cameras, these models were identified as optimal for evaluation due to their complementary strengths.

Several on-camera speed tests were performed on multiple cameras, and the runtime speed statistics were evaluated for YOLO v5 and SSD MobileNet V2. The average runtime speed statistics for various models were obtained from multiple on-time speed tests conducted on different cameras, as detailed in Table 3 below.

**Table 3.** Average Runtime Speed Statistics for Several On-time Speed Tests.

Model Name	Average Runtime Speed (s)		Final Average Runtime Speed (s)
	Camera 1	Camera 2	
YOLO v5	2.5	3.0	2.75
SSD MobileNet V2	0.2	0.2	0.20

One camera resulted in an average runtime speed of 2.5 s and 0.2 s for YOLO v5 and SSD MobileNet V2, respectively, while the other camera resulted in an average runtime speed of 3 s and 0.2 s for YOLO v5 and SSD MobileNet V2, respectively. The average was obtained from 5 runs recorded for each model to evaluate the speed. The final average runtime speed was recorded as 2.75 s for YOLO v5 and 0.2 s for SSD MobileNet V2, respectively. Based on these results, YOLO v5 was identified as having a relatively slower runtime speed than SSD MobileNet V2 on the camera. This significant difference in speed was crucial, as real-time performance is paramount for effective PPE detection in dynamic construction environments. Consequently, among these models, SSD MobileNet V2 was identified as having the best performance on the available hardware, and therefore, SSD MobileNet V2 was used for the backbone of the system.

SSD MobileNet V2 is a highly effective SSD architecture model for images and objects with a backbone network consisting of a MobileNet v2 feature extractor which has gained popularity for its lean network and novel depth-wise separable convolutions. It is a Deep Learning Convolutional Neural Network (DL-CNN) architecture containing the initial fully convolutional layer with 32 filters, followed by 19 residual bottleneck layers. SSD MobileNet V2 was expertly fine-tuned on a comprehensive PPE object detection dataset to perform well in detecting objects in bounding boxes on mobile devices with high accuracy performance. SSD MobileNetV2 is considered highly advantageous for deployment due to several key reasons. MobileNetV2 is specifically designed to be lightweight and efficient, making it well-suited for deployment on resource-constrained devices like mobile phones, edge devices, and IoT devices. Its architecture uses depth-wise separable convolutions and other techniques to reduce the number of parameters and computations without significantly sacrificing accuracy. SSD combined with MobileNetV2 architecture allows for fast object detection in real-time applications, which is crucial for applications where low latency is important, such as surveillance. MobileNetV2 strikes a good balance between model size, speed, and accuracy. While it may not achieve the absolute highest accuracy compared to larger and more complex models, its performance is often sufficient for many practical applications where speed and efficiency are prioritized over marginal gains in accuracy. The efficiency of MobileNetV2 enables deployment in various scenarios where computational resources are limited. This includes not only mobile devices but also embedded systems and other edge computing environments where power consumption and heat dissipation are critical concerns. MobileNetV2 is part of the TensorFlow Lite model repository and has widespread support in the deep learning community. There are numerous resources, pre-trained models, and optimization techniques available, making it easier for developers to deploy and fine-tune the model for specific applications. In summary, SSD MobileNetV2 is highly favored for deployment due to its compact size, efficient architecture, real-time performance capabilities, and suitability for resource-constrained environments. For these reasons, SSD MobileNet V2 was used as the backbone of the system.

SSD MobileNet V2 supports various input resolutions, allowing for flexibility in balancing between computational cost and detection accuracy based on the specific requirements of the deployment scenario. MobileNet V2's architecture is conducive to transfer learning, allowing pre-trained weights to be fine-tuned on specific datasets with relatively small amounts of new data, which accelerates the training process and enhances performance. The model has demonstrated robustness in detecting objects across a variety of conditions, including different lighting, weather, and occlusions, making it suitable for real-world applications where conditions are not always controlled [28]. Due to its popularity and support from the deep learning community, MobileNet V2 benefits from continuous improvements and widespread validation in various applications, ensuring a reliable and up-to-date architecture. The architecture is well-supported by many hardware platforms and development tools, facilitating seamless integration into existing systems and workflows, which simplifies deployment and maintenance. SSD MobileNet V2 is compatible with many edge AI platforms, such as NVIDIA Jetson, Google Coral, and various ARM-based devices, providing a wide range of deployment options for different hardware environments. MobileNet V2 can be easily quantized to reduce model size and increase inference speed without significant loss in accuracy, making it even more suitable for deployment on edge devices with strict resource constraints. The architecture can be extended to support multi-task learning, such as simultaneous object detection and segmentation, which can enhance the utility of the model in complex applications. The model is designed to be energy-efficient, which is crucial for battery-powered devices and applications where minimizing power consumption is essential to extend device operation time. The architecture can be scaled up or down depending on the performance requirements and resource availability, offering flexibility in deployment from small-scale projects to large-scale implementations. SSD MobileNet V2 benefits from various optimization tools and frameworks designed for edge AI, such as TensorFlow Lite, OpenVINO, and EdgeTPU, which streamline the process of deploying optimized models on edge devices. The architecture has been validated in numerous real-world applications, providing a track record of reliability and effectiveness in diverse and practical use cases, including autonomous vehicles, smart cameras, and mobile applications. These points further emphasize the suitability and advantages of using SSD MobileNet V2 as the backbone architecture for the object detection model, especially in the context of deployment on resource-constrained edge devices.

AI model training was performed iteratively as new data were labeled and validated for training. Existing and new data were split 70%/30% in the training and validation set, respectively. Non-Maximal Suppression (NMS) was employed to selectively retain only the precise bounding boxes and to enhance the accuracy of the model. While NMS is typically used during inference to filter overlapping bounding boxes, NMS was integrated into the training process to refine the bounding box predictions. The initial stage of NMS involved discarding previously anticipated bounding boxes with an identity probability significantly lower than the specified NMS threshold. In the present system, the NMS threshold was defined as 0.5, signifying that all predicted bounding boxes surpassing a detection probability of 0.5 were preserved within the system. The integration of NMS into the training process, although unconventional, was a deliberate strategy to enhance the model's performance. To ensure the transparency and explainability of the model, various measures were implemented, including meticulous tracking of label instructions, robust model versioning, and comprehensive data versioning. Various fairness-enhancing techniques, such as re-sampling and re-weighting, were implemented and tested to ensure the model performs equitably across diverse populations.

### 3.3. Testing

During training, the model performance was monitored periodically on the validation set by calculating COCO evaluation metrics. The COCO metrics are a standard set of metrics widely used in object detection tasks to assess model performance and include

measurements such as AP, which considers different Intersection over Union (IoU) thresholds to evaluate the accuracy of the predicted bounding boxes. The COCO evaluation metrics provide a comprehensive assessment by evaluating precision and recall across various IoU thresholds and object sizes, offering insights into how well the model performs in detecting objects under different conditions.

Overall model performance was measured over a subset of 794 images at the final stage. Due to the small sample size, an additional test set was omitted for AI demo creation. Tests with out-of-distribution data were conducted manually by a human-in-the-loop workflow to identify by sampling and ranking unlabeled data by AI model uncertainty. The sampling strategy was top–bottom, i.e., the 50 images with the lowest and highest AI model confidence, respectively, were sampled for manual labeling to increase information gain and reduce bias. The authors acknowledge that a production AI model will require a dataset for training, validation, and testing. For the demonstration, the best model was selected by monitoring the COCO scores on the validation set, and training was stopped if additional training did not show any improvements in the metrics for 90 minutes. Evaluation metrics are further elaborated below.

AP is a primary metric used for object detection tasks. AP combines precision and recall into a single value by summarizing the precision–recall curve and provides a single score that reflects the model’s performance. AP is calculated at different IoU thresholds (e.g., 0.50, 0.75) and averaged to provide a comprehensive evaluation of the model’s precision at different levels of overlap between predicted and ground truth bounding boxes. Precision measures the proportion of correctly identified positive instances out of all instances identified as positive, while recall measures the proportion of correctly identified positive instances out of all actual positive instances. AP reflects the model’s ability to accurately detect objects with minimal FPs and False Negatives (FNs).

Average Recall (AR) measures the ability of the model to find all relevant objects. It is calculated at various IoU thresholds and averaged over different object sizes to evaluate how well the model can recall objects of different scales. AR provides an indication of how well the model can identify objects regardless of the precision of their bounding boxes. Higher AR values suggest that the model is successful in locating the majority of objects in the images, even if the exact boundaries are not perfectly aligned.

IoU measures the overlap between the predicted bounding box and the ground truth bounding box. Higher IoU values indicate better alignment between the predicted and actual locations of objects.

By using these metrics, the model’s performance was rigorously assessed, ensuring that it could effectively detect PPE in real-world scenarios. The manual labeling process for out-of-distribution data further enhanced the model’s robustness, addressing potential biases and improving its generalization capabilities. All metrics were calculated based on the top 100 highest-scoring detections per image across all categories. The evaluation metrics for bounding box detection are the same in every aspect except for the IoU computation, which is performed over bounding boxes for detection. In summary, the testing process combined automated evaluation using COCO metrics with manual verification through a human-in-the-loop approach. This comprehensive testing strategy ensured that the model selected for the demonstration was both accurate and reliable, capable of performing effectively in diverse and challenging environments.

### 3.4. Deployment on Edge

CVTool [36] was used to efficiently map the SSD MobileNet V2 network trained with an industry-standard tool (TensorFlow 1.10) to run on Ambarella processors. The iPRO cameras utilize the Ambarella-built CV SOC, enabling real-time complex data analytics, delivering superior image quality, and optimizing vital system resources such as power and network bandwidth. The CNN trained using industry-standard tools like Caffe, TensorFlow, and PyTorch can seamlessly map onto Ambarella processors, leveraging the efficiency of CVflow. The chip architecture of CVflow has been meticulously designed with a profound comprehension of fundamental computer vision algorithms. Distinguished from general-purpose CPUs and GPUs, CVflow encompasses a dedicated vision processing engine that operates based on a high-level algorithm description. This unique architecture facilitates scaling performance to trillions of operations per second while simultaneously maintaining remarkably low power consumption levels. CVTool is compression efficient and reduces bandwidth and storage costs. By leveraging CVTool, it is possible to build a fast, intelligent, and powerful system capable of efficiently executing cutting-edge neural networks. This is achieved with minimal exertion, ensuring high precision and low power consumption. Notably, the integration of compression technologies significantly diminishes the size of transmitted data, resulting in reduced storage costs and bandwidth usage. An exemplary characteristic of Ambarella processors lies in their comprehensive image processing pipeline, encompassing a range of functionalities, including HDR, EIS, and dewarping. This state-of-the-art Image Signal Processor (ISP) guarantees that each frame of every video attains a level of visual excellence, thereby augmenting safety and enriching the viewing experience.

Due to limited computing resources and proprietary software to run the models on the edge, the search space was limited by supported ops. However, SSD MobileNet V2 was found to meet the requirements. After training, the checkpoint and model were converted into a TensorFlow 1 Frozen Graph. The frozen graph was then passed into CVTool, which sparsifies the model using post-training quantization (into INT8 and INT16) and pruning. The trained neural network was analyzed and optimized using the CNN Generation tool for hardware via sparsification and quantization. Subsequently, the optimized network underwent the compilation process using a dedicated tool, generating a program written in a high-level language instead of generic low-level operators. This program was tailored for execution on CVflow hardware. The compiled program, referred to as a DAG executable binary, was then transmitted to the CVflow hardware for execution. The resulting output was rigorously validated, and if necessary, the network underwent retraining. The input file was 17.5 MB, and the output files consisted of two binaries: one for the model and one for the anchor boxes totaling 4.4 MB, which was nearly a four times reduction in file size. The model was then shipped in the form of an AdamApp, which includes the pre- and post-processing code.

The technology has been specifically developed for deployment on embedded platforms on the edge, such as battery-powered cameras, where power consumption emerges as a critical consideration. Ambarella processors consistently exhibit superior power efficiency compared to competing solutions, often surpassing them by a factor of 5 or more while simultaneously delivering equivalent or superior outcomes.

The AdamApp has a logic to trigger alerts once per non-compliance instance based on the number of class predictions per timespan the instances must land in, thus avoiding sending out too many alerts to the destination about the same non-compliance incident. If non-compliance predictions in the last  $x$  seconds were more than  $y$ , and if no alerts were sent in the last  $x$  seconds, an alert is sent. Here,  $x$  and  $y$  are configurable parameters depending on AI model FPS, accuracy in the deployed environment, and clients' preference to be over-alerted or under-alerted. Sensible defaults for  $x$  and  $y$  might be 2 and 4 (assuming the model is running at least 15 FPS), respectively.

Accordingly, the proposed system leverages AI-enabled edge-capable IP security cameras, specifically iPRO (formerly Panasonic) devices, deployed at construction sites

for enhanced safety surveillance. These cameras are equipped to perform real-time processing locally on the edge without reliance on a centralized VMS. By embedding AI processing capabilities directly within the cameras, the system autonomously detects instances of non-compliance with safety protocols, such as workers not wearing required hard hats and safety vests. The AI-enabled edge processing in the iPRO cameras significantly enhances the system's reliability and responsiveness, which is crucial for real-time PPE detection and alerting. The embedded AI capabilities allow the cameras to process video feeds and detect safety violations independently, eliminating the need for continuous high-speed internet connectivity and extensive cloud-based processing. This decentralized approach not only reduces latency but also improves the overall scalability and efficiency of the safety surveillance system. Using advanced computer vision techniques based on SSD MobileNet V2, for instance, segmentation and classification, the system identifies and labels individuals within video frames who are not adhering to safety regulations. This approach not only ensures immediate detection and response but also reduces latency and dependency on high-speed internet typically required by cloud-based systems. Additionally, the iPRO cameras are designed with network and digital I/O outputs that facilitate the connection of external sensors, such as sirens or warning lights, allowing for immediate on-site alerts in case of safety violations. This real-time processing ability ensures that safety violations are identified and addressed instantaneously, minimizing the time gap between detection and corrective action. By functioning independently of a VMS, the system reduces the need for extensive infrastructure and lowers operational costs, making it a cost-effective solution for safety monitoring. The cameras' edge processing capabilities also minimize data transmission to central servers, enhancing data privacy and reducing bandwidth usage, thus ensuring an efficient and comprehensive safety surveillance system suitable for various construction site environments. Overall, the integration of these advanced, AI-enabled cameras provides a robust, reliable, and scalable solution for maintaining high safety standards on construction sites.

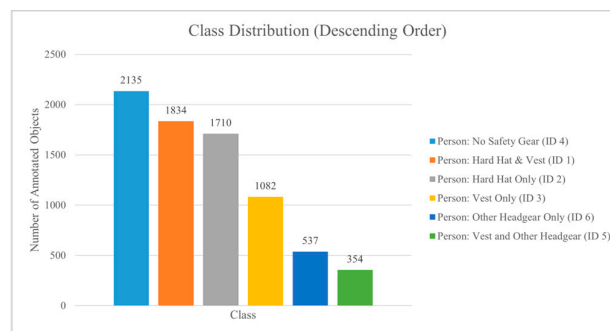
During the deployment of the model on edge devices, several challenges were encountered and effectively addressed to ensure optimal performance and functionality. Edge devices typically have constrained computational capabilities compared to cloud servers or desktops. This limitation necessitated optimizations in model architecture and processing techniques to ensure that the deployed model could operate efficiently within these constraints. The proprietary nature of the software environment on edge devices posed challenges in terms of compatibility and supported operations. This required careful selection and adaptation of tools and frameworks to ensure seamless integration and functionality. However, SSD MobileNet V2 proved suitable for meeting the requirements. Leveraging hardware acceleration, such as Ambarella processors with CVflow technology, required specialized knowledge and tools to effectively map and optimize the model. Techniques like post-training quantization and model pruning were employed to reduce computational complexity and enhance execution speed without compromising accuracy. Ensuring that the model and associated data could be efficiently transmitted and deployed on edge devices with limited storage and bandwidth was critical. Compression techniques and optimization strategies were implemented to minimize file sizes and reduce transmission overhead. Meeting real-time performance requirements for safety surveillance applications involved fine-tuning model inference speeds and optimizing data processing pipelines. This included optimizing algorithms and hardware configurations to achieve low-latency responses essential for timely detection and alerting. Integrating the optimized model into the existing edge device ecosystem required rigorous testing and validation to ensure compatibility, reliability, and accuracy under real-world conditions. This involved iterative adjustments and validations to fine-tune the model's performance and address any discrepancies. Designing the deployment framework to be scalable across multiple edge devices while maintaining ease of maintenance and updates posed additional challenges. Strategies for version control, remote management, and scalability planning were essential to support ongoing operations and future expansions. In summary,



overcoming these challenges involved a combination of advanced technical solutions, strategic optimizations, and meticulous testing to ensure the effective deployment and operation of our AI model on edge devices for enhanced safety surveillance applications.

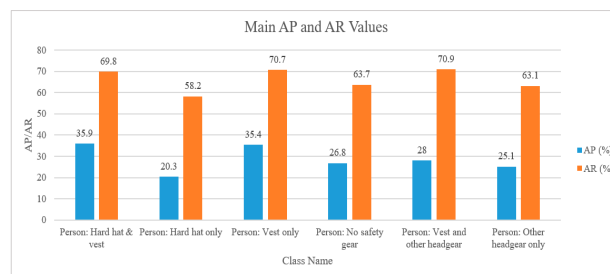
#### 4. Experiments and Results

The AI model was implemented, trained, and tested using TensorFlow 1.15.5 and CUDA 10.0. TensorFlow Object Detection API was used to train the models. For backward compatibility, the final model was exported using TensorFlow 1.10.1–1.12.0 (Google LLC, Mountain View, CA, USA) and NVIDIA CUDA 9.2 (NVIDIA, Santa Clara, CA, USA) environment. The overall performance of the model was evaluated using several metrics such as IoU, AP, and AR values class-wise, mAP and Mean Average Recall (mAR), the standard metrics for evaluating computer vision classifiers based on factors such as the imbalance of the dataset. AP and mAP values were used to cater to the requirement of quantifying both classification and localization performance simultaneously. Predictions were obtained at the 50% threshold. The class distribution of the dataset is shown in Figure 4 below.



**Figure 4.** Class distribution.

As seen from the graph above, the dataset is more biased toward the person: no safety gear class (ID 4) and less towards the person: vest and other headgear class (ID 5). Figure 5 below shows the class-wise AP and AR values obtained by the model for COCO evaluation metrics at the iteration of 373,051.



**Figure 5.** Main AP and AR values for each class.

It can be observed that the AR value is greater than the AP value for all the classes. This highlights a specific aspect of the model's performance. The model is effective at identifying and locating objects within the images, which results in high recall values. It can detect the presence of objects consistently, ensuring that most objects are found, even if the predicted bounding boxes are not perfectly accurate. The higher AR also indicates that the model has a strong capability in detecting objects, ensuring that few objects are missed. The model demonstrated the following AP and AR values for COCO evaluation metrics for different IoU values:

AP @ [IoU=0.50 | area=all | maxDets=100]: 0.491;  
 AP @ [IoU=0.50:0.95 | area=all | maxDets=100]: 0.286;  
 AP @ [IoU=0.50:0.95 | area=large | maxDets=100]: 0.438;  
 AP @ [IoU=0.50:0.95 | area=small | maxDets=100]: 0.051;  
 AP @ [IoU=0.50:0.95 | area=medium | maxDets=100]: 0.221;  
 AP @ [IoU=0.75 | area=all | maxDets=100]: 0.295;  
 AR @ [IoU=0.50:0.95 | area=all | maxDets=1]: 0.3;  
 AR @ [IoU=0.50:0.95 | area=all | maxDets=10]: 0.464;  
 AR @ [IoU=0.50:0.95 | area=all | maxDets=100]: 0.476;  
 AR @ [IoU=0.50:0.95 | area=large | maxDets=100]: 0.661;  
 AR @ [IoU=0.50:0.95 | area=small | maxDets=100]: 0.087;  
 AR @ [IoU=0.50:0.95 | area=medium | maxDets=100]: 0.409.

The classification loss was recorded as 13.33, and the localization loss was recorded as 0.95. Figure 6 below shows some of the prediction results obtained by the system for some images from the Hard Hat Workers Dataset [34] and images extracted from videos on the internet.





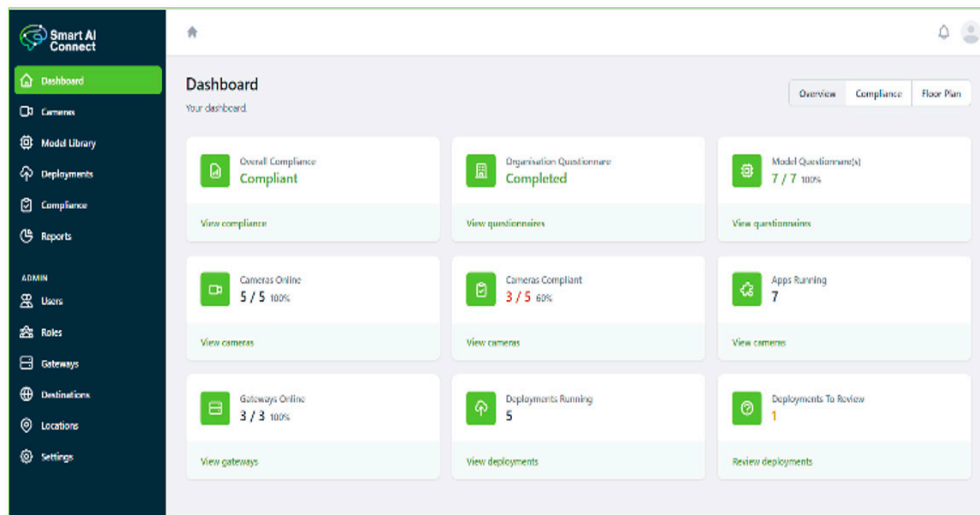
**Figure 6.** Sample prediction results of selected classes, corresponding GT images, and original images. [34]

Based on the findings depicted in Figure 6, the prediction images exhibit impeccable segmentation outcomes in comparison to the Ground Truth (GT) images. The observed misclassification ratio for each segmented class was extremely low, thereby greatly facilitating the decision-making process for the target class and effectively eliminating the misclassification challenge within the segmentation network. Figure 7 below shows some of the prediction results obtained by an iPRO camera mounted at eye level at a private apartment in Queensland, Australia.



**Figure 7.** Sample prediction results obtained by an iPRO camera on real-world images for hard hat and vest class, vest-only class, and hard hat-only class with their prediction scores.

All the predictions in Figure 7 have an approximate 100% confidence and precise bounding boxes covering the person in front of the CCTV camera ensuring the safety of the person. Figure 8 below shows some of the screenshots of the management platform, which is visible to the safety officer at the destination setup in the gateway.



(a) Dashboard with comprehensive options for camera statuses, model library, deployments, compliance, reports, users, gateways, and locations.

Smart AI Connect Cameras List:

Name	Site	Gateway	Status	IP Address	Make	Camera Model	AI Model	Compliance
AiHub 360	Queensland AI Hub	Queensland AI Hub	Online	192.168.2.51	I-PRO	WV-S4176		Compliant
Inside Demo - No AI Allowed	The Lake	Demo NUC	Online	192.168.2.12	I-PRO	WV-S2236L		Warning
Lake Bullet	The Lake	Demo NUC	Online	192.168.2.11	I-PRO	WV-S1536L		NonCompliant
Test Dome on Stand	SmartAI Headquarters	Cohort Test NUC	Online	192.168.1.3	I-PRO	WV-S2236L		Compliant
Test Multilens on Stand - On Loan	SmartAI Headquarters	Cohort Test NUC	Online	192.168.1.2	I-PRO	WV-S8544L		Compliant
WP Street USA	WP, USA	USA Demo NUC	Online	192.168.151.217	I-PRO	WV-S1536L		Compliant

(b) Camera Status

**Figure 8.** Screenshots of the management platform.

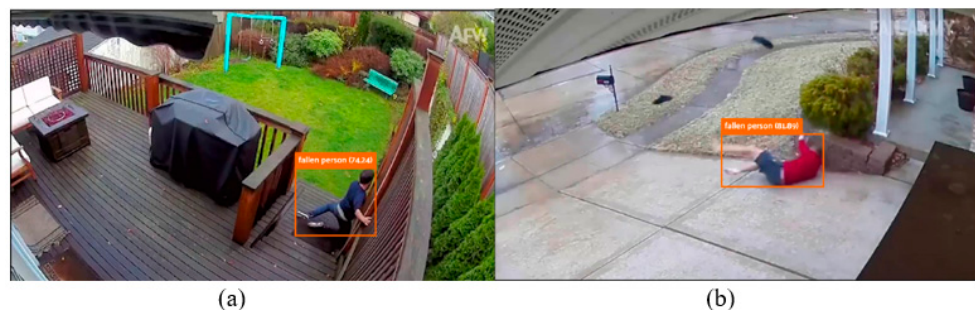
The proposed solution was extended to detect slips, trips, and falls. Slips are usually caused by wet or slippery surfaces or spilled items, often resulting in backward falls. Trips occur when an obstacle causes a person to stumble, typically leading to forward falls. Falls commonly result from either slipping or tripping incidents. Accordingly, the model was fine-tuned to recognize seven classes of objects related to various stages of losing balance and falling. These classes include “Falling forward”, “Falling backward”, “Fallen forward”, “Fallen backward”, “Lending a helping hand”, “Standing”, and “Other”. The “Lending a helping hand” class refers to situations where an individual assists another person who is experiencing or recovering from a slip, trip, or fall. This class captures scenarios where one person is actively helping another to regain balance, get up from a fall, or prevent a fall from occurring. The slip, trip, and fall detection model was designed for use in environments where such accidents are a risk, such as nursing homes, hospitals, and public spaces, and aims to promptly detect and alert caretakers, staff, or emergency services during a slip, trip, or fall, potentially reducing the risk of injury through rapid response. The dataset consisted of images extracted from internet videos, primarily surveillance camera fail compilations and real-world instances of falling captured by a mobile phone camera in an office environment in Brisbane, Australia. The data were collected



from diverse sources and meticulously curated to avoid any bias. A team of human annotators reviewed and labeled the data to ensure accuracy and consistency. For testing, 181 samples were used. The model was trained to identify slip and fall incidents irrespective of an individual's age, gender, race, or any other characteristic. It demonstrated the following AP and AR values based on COCO evaluation metrics for different Intersection over Union (IoU) values:

AP @ [IoU=0.50 | area= all]: 0.499;  
 AP @ [IoU=0.50:0.95 | area= all]: 0.325;  
 AP @ [IoU=0.50:0.95 | area= large]: 0.454;  
 AP @ [IoU=0.50:0.95 | area= small]: 0.377;  
 AP @ [IoU=0.50:0.95 | area=medium]: 0.342;  
 AP @ [IoU=0.75 | area= all]: 0.352;  
 AR @ [IoU=0.50:0.95 | area= all | maxDets=1]: 0.423;  
 AR @ [IoU=0.50:0.95 | area= all | maxDets=10]: 0.496;  
 AR @ [IoU=0.50:0.95 | area= all | maxDets=100]: 0.501;  
 AR @ [IoU=0.50:0.95 | area= large | maxDets=100]: 0.628;  
 AR @ [IoU=0.50:0.95 | area= small | maxDets=100]: 0.448;  
 AR @ [IoU=0.50:0.95 | area=medium | maxDets=100]: 0.489.

The results indicate that the model performs reasonably well in detecting slips, trips, and falls, especially considering the varying object sizes and IoU thresholds. The model shows higher precision and recall for larger objects compared to small and medium ones. This suggests that the model is more accurate and consistent in detecting larger instances of slips, trips, and falls but still performs well for smaller and medium-sized instances. The metrics highlight the model's strengths and areas for improvement, such as enhancing detection accuracy for smaller objects and increasing precision at higher IoU thresholds. Figure 9a [37] and Figure 9b [38] are sourced from images extracted from internet videos, specifically from surveillance camera failure compilations.



**Figure 9.** Sample prediction results obtained for fallen forward (a) [37] and fallen backward (b) [38] classes.

The prediction includes a bounding box around the detected person, demonstrating the model's capability to accurately classify and localize forward and backward fallen incidents. These samples highlight the model's effectiveness in detecting different types of falls, showcasing its potential utility in environments where monitoring and rapid response to slip, trip, and fall incidents are crucial.

## 5. Result Analysis and Discussion

According to the graph in Figure 5, the mAP of the model is 29, and the mAR of the model is 66. The model has a high mAR but a low mAP. The model demonstrates accurate classification of the majority of positive samples; however, it exhibits a notable number of false positives (FPs), indicating the misclassification of negative samples as positive. This can be improved by training the model on more data that include more objects with use

cases limiting the model performance and by techniques such as active learning and AI uncertainty sampling.

The model has the following limitations, as well. The system exhibited performance issues under particular circumstances. The processing time for a  $400 \times 400$  pixels image will be approximately 4 times that of a  $200 \times 200$  pixels image (image resolution vs. latency). Latency increases proportionally with image pixel count. The model performed less than optimally in this situation. The model was less accurate under low lighting conditions or when the air was humid. While the dataset contained images taken at night, these images were less well-present and usually had poor lighting. The model was only trained on images in full color and black and white and not on infrared or multi-spectrum. A person detector that undergoes training using partial annotations has the potential to acquire the ability to emulate the non-exhaustive patterns present in the training data. Consequently, the detector trained under these circumstances may exhibit reduced efficacy in accurately detecting individuals belonging to specific subpopulations within certain contextual scenarios. Images downloaded from the internet revealed a significant imbalance, with a disproportionate number of images originating from China. Notably, skin tone emerges as a crucial attribute to be considered for ensuring fairness, yet this dataset lacks representation in that regard. However, the training and testing data contained a variety of skin tones and genders to minimize biases. Measures taken to ensure the model is not perpetuating existing biases include using diverse training data, evaluating the model's performance on different subgroups of the data, and conducting ongoing monitoring of the model's performance in the field.

The majority of the images in the dataset comprised images downloaded from the internet. The dataset had fewer images captured from the surveillance camera, which was also a limitation of the model. Different types of headgear that share one or multiple similar characteristics (e.g., shape, color) were present in the evaluation data instances, which degraded the model's performance. The model was shown to have low performance in distinguishing hard hats and other types of headgear, such as hoodies, caps, or beanies. Some caps, beanies, and hard hats almost looked similar in shape and color. Sometimes, the most significant difference between them was the material they were made of. Therefore, the model had difficulty distinguishing hard hats with different types of headgear (hoodies, caps, beanies, etc.).

On some occasions, hair was identified as headgear (hard hat/hoodie/cap/beanie) and vice versa. Sometimes, a person holding a vest was identified as a vest-wearing person, and a person holding a cap/beanie was identified as a cap/beanie-wearing person. Furthermore, a person holding a hard hat was identified as a hard hat-wearing person on some occasions. Sometimes, persons were misclassified with some objects in the background with similar colors and shapes (e.g., chairs, pictures on screens, advertisements, wall hangings, etc.), and headgear was misclassified with some objects in the background with similar colors and shapes. The headgear may not be picked when the headgear and the background behind the headgear are similar. Bald heads were identified as hard hats in some cases. Objects with luminant stripes (e.g., cups, safety belts, clothes) were misclassified as vests.

Apart from the above limitations, the model may have the following limitations as well. People who are distant from the camera (a pupillary distance of  $<10$  px) and people who are too small might not be detected when the image resolution gets smaller. The performance will degrade when running on CCTV streams zoomed out too far making people hard to detect. The model is not designed to estimate the size of a crowd. Therefore, when the crowd in front of the camera increases, the predictions will be incorrect. Vests, hard hats, hoodies, caps, and beanies might not be detected in different postures and orientations of people. The model needs visible landmarks of luminant stripes for vests to be identified correctly. When people are looking away from the camera (pan  $>90^\circ$ , roll  $>45^\circ$ , or tilt  $>45^\circ$ ), hard hats, hoodies, caps, and beanies might not be detected properly. Partially hidden, obstructed people and blurry people might not be detected. When people are

overlapping, the models' performance might degrade. The model's performance may be impacted by variations in weather, camera angle, and other environmental factors that are not present in the training data. This means that the model may not perform as well in situations that differ significantly from the conditions it was trained on.

## 6. Comparative Analysis

Speed, mAP, and mAR of the proposed system and some state-of-the-art methods are compared, as shown in Table 4 below.

**Table 4.** Comparative Analysis.

Model Name	Speed on			Classes	Parameters
	CPU (ms) for 37 FPS	mAP (%)	mAR (%)		
YOLOv5	28	50.5	44.1	80	7.30 million
Faster R-CNN (ResNet-50)	85	36.8	34.0	80	41.50 million
EfficientDet-D0	50	33.6	32.5	90	3.90 million
YOLOv3	45	33.0	31.4	80	62.00 million
SSD ResNet-50 FPN	80	39.0	35.0	90	44.00 million
Baseline: SSD MobileNet V2 (COCO) **	27	22.0	27.0	90	3.47 million
Proposed method	27	29.0	66.0	10	3.47 million

\*\* TensorFlow Model Zoo reference metric is "DetectionBoxes\_Precision/mAP" (aka AP @[IoU=0.50:0.95 | area= all | maxDets=100 ]").

Based on the performance metrics provided in Table 4, it is evident that the proposed method outperforms several well-known object detection models. The proposed method matches the speed of SSD MobileNet V2 (COCO) with a runtime of 27 ms, which is significantly faster than other models, such as Faster R-CNN (85 ms) and SSD ResNet-50 FPN (80 ms). The proposed method achieves a mAP of 29%, which represents an increase of 7% compared to the baseline SSD MobileNet V2 (COCO). This improvement places it ahead of EfficientDet-D0 (33.6%), YOLOv3 (33.0%), and even the high-parameter YOLOv5 (50.5%). The mAR of the proposed method is 66%, which is substantially higher than all other models, including the baseline SSD MobileNet V2 (27%) and YOLOv5 (44.1%). This indicates that the proposed method is more effective in identifying relevant instances. The proposed method maintains a low parameter count of 3.47 million, similar to SSD MobileNet V2, making it efficient in terms of computational requirements while achieving superior performance metrics. In conclusion, the proposed method demonstrates a significant improvement in detection accuracy (mAP and mAR) while maintaining efficient computational performance, making it a superior choice compared to the existing models listed in the table. The proposed method's balanced approach of achieving high accuracy and recall while retaining efficiency clearly sets it apart as a superior alternative.

The dataset was extended by including a larger and more diverse set of images directly from real-world construction sites, capturing various lighting conditions, weather scenarios, and different types of PPE. The representation of diverse demographic groups was increased to ensure fairness and reduce potential biases in the model. The model was trained again with 3 new main classes: 'Person: Cap/beanie', 'Person: Vest and cap/beanie', and 'Object: Hard hat'. The 'Person: Cap/beanie' class represents the people wearing headgear other than a hard hat, e.g., cap, beanie, hoodie, hat, etc., and wearing no hard hat and no high visible safety vest with reflective stripes. 'Person: Vest and cap/beanie' represents people wearing only a high visibility safety vest with reflective stripes and headgear other than a hard hat, e.g., cap, beanie, hoodie, hat, etc. The 'Object: Hard hat' class represents

hard hat objects, either worn or not. Including this class ensures that the model can accurately detect various stages of wearing and removing the headgear, thereby enhancing the model's accuracy and robustness. Each class was assigned a distinct RGB color based on a traffic light color schema facilitating visual aid during the data labeling process.

The data were collected over a 7-day period by three iPRO cameras placed at actual construction sites, both indoors and outdoors. Data collection occurred during the morning, afternoon, evening, and night, for at least 2 h each session, and included images of different lighting conditions, weather scenarios, and different types of PPE. Each primary class appeared with at least three different subjects. Based on these camera tests, a constraints write-up was created to detail minimum and recommended camera resolution sizes, angle ranges, simultaneous detection counts, minimum illumination levels, maximum number of cameras, and watch list limits. Using this information, the model's performance was further improved through techniques such as active learning.

Data augmentation techniques such as barrel distortion to mimic the fisheye of the cameras, mosaic to make objects smaller, random cropping, flipping, rotation, scaling, and color jittering to increase the variability of training samples were applied. Barrel distortion mimics the fisheye effect of wide-angle cameras. This distortion makes images appear as if they were captured with a fisheye lens, introducing curvature. It helps the model learn to recognize objects even when they are distorted. Mosaic makes objects appear smaller and increases the number of objects in a single image. By combining multiple images into a single image, the model is exposed to a greater number of object instances, which helps in learning to detect smaller objects and improves robustness in crowded scenes. Random cropping introduces variability in object positioning. Portions of the image were randomly cropped, which helps the model learn to detect objects regardless of their position within the frame. Images were flipped horizontally or vertically, providing variation in orientation. This technique helps the model recognize objects from different angles and orientations. Rotation simulates different angles of view. Images were rotated by random degrees, and this helped the model learn to detect objects even when they were tilted or viewed from an angle. Images or objects within images were resized to help the model generalize to objects of various sizes, ensuring accurate detection irrespective of object scale. Color jittering introduces variability in color representation. Random changes were made to the brightness, contrast, saturation, and hue of the images. This helps the model become invariant to lighting conditions and color variations, improving performance under different lighting scenarios.

The test dataset was extended by including samples from videos captured by a regular surveillance camera (iPRO camera) placed indoors pointing to an outdoor construction site in Brisbane, Australia (200 samples), videos captured by an iPhone 11 Pro Max mobile phone camera from an outdoor construction site in Brisbane, Australia (25 samples), videos recorded by an iPRO camera placed in an indoor construction site in Queensland, Australia (163 samples), videos captured by an iPhone 11 Pro Max mobile phone camera from an indoor construction site in Brisbane, Australia (140 samples), and frames extracted from relevant sections of YouTube videos hashed using md5 algorithm (389 samples). The training and testing data contained a variety of skin tones and genders to minimize biases. The results are listed in Table 5.

The F-score is the harmonic mean of precision and recall, providing a single metric that balances both. An IoU of 0.50 means that the predicted bounding box must overlap at least 50% with the ground truth. The F-score, when considering a 50% overlap between predicted and actual bounding boxes, is 52%, which suggests a moderate level of precision and recall. It is reasonably good at correctly identifying and localizing objects when a 50% overlap is considered. At an IoU threshold of 0.50, the mAP is 43%. This indicates that, on average, the precision of the model's predictions is fairly moderate. mAP over multiple IoU thresholds averages the precision across multiple IoU thresholds, typically from 0.50 to 0.95 in steps of 0.05. The mAP across a range of IoU thresholds is 18%. This lower value suggests that the model's precision decreases as the required overlap between predicted



and ground truth bounding boxes increases. The mAR across IoU thresholds from 0.50 to 0.95 is 25%. This indicates that the model has a moderate ability to correctly identify all relevant instances, but performance varies with different levels of overlap. The mAP value drops to 18% when averaged over IoU thresholds from 0.50 to 0.95. This suggests that the model struggles more as the overlap requirement becomes stricter, indicating potential issues with precise localization. An mAR of 25% also indicates that while the model can detect relevant objects, its recall varies significantly with different IoU thresholds. The test dataset includes a variety of sources, such as different camera types (iPRO cameras and iPhone 11 Pro Max) and locations (indoor and outdoor construction sites), as well as frames from YouTube videos. This diversity likely helps in evaluating the model's robustness across different conditions. The inclusion of various skin tones and genders in the training and testing data is aimed at reducing biases, ensuring the model performs equitably across diverse populations. The results suggest that while the model performs moderately well in terms of precision and recall at a basic IoU threshold, there is room for improvement, especially in achieving higher precision and recall at stricter IoU thresholds.

**Table 5.** The Results after Extending the Dataset.

Metric	Value (%)
F-score @[ IoU=0.50 ]	52
mAP @[ IoU=0.50 ]	43
mAP @[ IoU=0.50:0.95 ]	18
mAR @[ IoU=0.50:0.95 ]	25

## 7. Conclusions and Future Work

The proposed approach can be used to identify people wearing hard hats and vests on AI-enabled CCTV cameras with or without VMS. The model could be used in a variety of applications, such as safety inspections in construction sites, factories, or other industrial settings, quality control in manufacturing or production lines, and surveillance and security systems in public places or private facilities. The model can be used as a tool to assist human operators in identifying potential safety hazards or non-compliance with PPE regulations. However, the model is not intended to replace human judgment or expertise, and its results should be carefully reviewed by qualified personnel before taking any action. It is important to carefully evaluate the model performance on a representative set of test data and consider its limitations and potential biases before deploying it in a real-world application. It is also important to ensure that appropriate safety measures are in place to mitigate potential risks associated with the use of the model.

The following are recommended to enhance the model's performance in the future. Even though the system demonstrates a considerable amount of detection accuracy and fast detection speed during practical implementations, workers located at a considerable distance from the fixed camera might not be adequately captured during the inspection process. In subsequent endeavors, significant attention will be devoted to the optimization of small target recognition with the objective of enhancing detection accuracy. It is expected to evaluate the model on open source benchmarks and some other real-world data captured from different environments and populations sliced by different image and people characteristics separately, represent the model's performance graphically (precision-recall curve), train the model again using images directly captured from a surveillance camera, increase the size of the input images, include more images by changing the position of holding headgear (ex: different stages of wearing and removing the headgear) into the training pool and extend the dataset for other safety equipment such as gloves, work boots, goggles, masks, respirators, etc. Future updates will include mechanisms to detect and account for unavoidable or fall activities in the workplace, ensuring comprehensive

action detection. Techniques such as adversarial debiasing, which offers a sophisticated approach to reducing bias by focusing on learning fair representations, will be implemented in future work to further enhance the fairness of the model across diverse populations. Infrared or multi-spectrum images will be used to further enhance the model's performance in low-light and adverse weather conditions.

**Author Contributions:** P.S. was the major contributor to writing the manuscript. P.S. engaged in the literature review, dataset curating, data labeling, model testing, analysis, and interpretation of the data. G.N. developed the LabelEngine for data labeling and labeled the data. G.N. extracted image data from videos and prepared the dataset by applying several augmentation techniques. G.N. trained, validated, tested, and evaluated the model and contributed to writing the manuscript. J.F. developed the AI management platform for visualizing the alerts being sent to the destination setup in the gateway with a dashboard, a model library, and reports on compliance, ethics, and governance while editing the manuscript. N.C. contributed to dataset curating, labeling, and testing data, meanwhile contributing to editing the manuscript. S.P. contributed to integrating the trained model into AdamApp while editing the paper. A.A. contributed to editing the paper. C.L. contributed to designing the edge architecture, including the fusion of iPRO AI edge-capable IP security cameras, AI models, and VMS, and editing the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was internally funded by Smart AI Connect. There were no research grants associated with the project. No organizations, stocks, or shares in companies (including holdings of spouse and/or children) would be affected financially by the publication of this manuscript. No consultation fees or other forms of remuneration from organizations may gain or lose financially, while none of the patents or patent applications' value may be affected by the publication of this manuscript.

**Institutional Review Board Statement:** All subjects gave their informed consent for inclusion before they participated in the study. Ethics approval is not required for this type of study. The study was conducted following the Australia's AI Ethics Principles: <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles> (accessed on 19 June 2024) and the Australian Government Privacy Act 1988: <https://www.oaic.gov.au/privacy/privacy-legislation/the-privacy-act> (accessed on 19 June 2024).

**Data Availability Statement:** Material and Code: The image/video data and the model files have been securely stored in AWS S3 with strict access management. All the data supporting the results reported in the manuscript, including codes, have been saved inside Smart AI Connect's private SharePoint folder. Annotators were all validated and trusted. The data can be made available when individual privacy could be compromised.

**Acknowledgments:** This application was developed by Smart AI Connect in collaboration with MaxusAI and iPRO for the Australian National Security Exhibition & Conference, held in Sydney in 2022. Noor Naddour and Rukshan Gunasena contributed to dataset annotation, while Benjamin Gonzalez contributed to data labeling and developing the MaxusAI Label Engine. Matthew Lee contributed to implementing the models on the iPRO camera (AdamApp) and conducting on-camera speed tests to compare the runtime speed of state-of-the-art techniques.

**Conflicts of Interest:** This research was internally funded by Smart AI Connect. J.F., N.C., S.P., and C.L. served as the CTO, Business Analyst, Computer Vision and MLOps Engineer, and CEO of Smart AI Connect, who received a salary from the company. G.N. was the Founder of MaxusAI, and A.A. was a lecturer at RMIT University. Author P.S. was an unpaid intern at Smart AI Connect.

## References

1. FACTS ON Safety at Work, ILO, Geneva, Switzerland, Apr. 2005. Available online: [https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms\\_067574.pdf](https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_067574.pdf) (accessed on 23 September 2022).
2. Construction Industry Profile, Safe Work Australia, Australia. Available online: <https://www.safeworkaustralia.gov.au/system/files/documents/1702/construction-industry-profile.pdf> (accessed on 23 September 2022).
3. Safety Helmets on Construction Sites. Available online: <https://www.commerce.wa.gov.au/publications/safety-helmets-construction-sites> (accessed on 23 September 2022).

4. Australian Work Health and Safety Strategy 2012–2022, Safe Work Australia, Canberra, Australia. Available online: <https://www.safeworkaustralia.gov.au/system/files/documents/1902/australian-work-health-safety-strategy-2012-2022v2.pdf> (accessed on 23 September 2022).
5. Australian Standard AS/NZS 1800:1998, Occupational Protective Helmets, Standards Australia, 1998. Available online: <https://store.standards.org.au/product/as-nzs-1800-1998> (accessed on 26 July 2024).
6. Australian Standard AS/NZS 4602:1999, High Visibility Safety Garments, Standards Australia, 1999. Available online: <https://store.standards.org.au/product/as-nzs-4602-1999> (accessed on 26 July 2024).
7. Kelm, A.; Laußat, L.; Meins-Becker, A.; Platz, D.; Khazaei, M.J.; Costin, A.M.; Helmus, M.; Teizer, J. Mobile passive radio frequency identification (RFID) portal for automated and rapid control of personal protective equipment (PPE) on construction sites. *Automat. Constr.* **2013**, *36*, 38–52.
8. Han, S.; Achar, M.; Lee, S.; Peña-Mora, F. Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring. *Vis. Eng.* **2013**, *1*, 6.
9. Elhanashi, A.; Dini, P.; Saponara, S.; Zheng, Q. Integration of deep learning into the IoT: A survey of techniques and challenges for real-world applications. *Electronics* **2023**, *12*, 4925. <https://doi.org/10.3390/electronics12244925>.
10. Mneymneh, B.E.; Abbas, M.; Khoury, H. Automated hardhat detection for construction safety applications. *Procedia Eng.* **2017**, *196*, 895–902.
11. Man-Woo, P.; Nehad, E.; Zhenhua, Z. Hardhat-wearing detection for enhancing on-site safety of construction workers. *J. Constr. Eng. Manag.* **2015**, *141*, 04015024.
12. Barnich, O.; Droogenbroeck, M.V. ViBe: A powerful random technique to estimate the background in video sequences. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 945–948.
13. Li, K.; Zhao, X.; Bian, J.; Tan, M. Automatic safety helmet wearing detection. *arXiv* **2018**, arXiv:1802.00264.
14. Wójcik, B.; Źarski, M.; Książek, K.; Miszczak, J.A.; Skibniewski, M.J. Hard hat wearing detection based on head keypoint localization. *arXiv* **2021**, arXiv:2106.10944.
15. Shrestha, K.; Shrestha, P.P.; Bajracharya, D.; Yfantis, E.A. Hard-hat detection for construction safety visualization. *J. Constr. Eng.* **2015**, *16*, 1–8.
16. Al Daghan, A.T.A.; Vineeta; Kesh, S.; Manek, A.S. A deep learning model for detecting PPE to minimize risk at construction sites. In Proceedings of the International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 9–11 July 2021; pp. 1–6.
17. Kamal, R.; Chemmanur, A.J.; Jose, B.A.; Mathews, S.; Varghese, E. Construction safety surveillance using machine learning. In Proceedings of the International Symposium on Networks, Computers and Communications (ISNCC), Montreal, QC, Canada, 20–22 October 2020; pp. 1–6.
18. Yu, K.; Liu, H.; Li, T.; Liu, X.; Yao, L.; Yan, Y. A protective equipment detection algorithm fused with apparel check in electricity construction. In Proceedings of the 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 22–24 May 2021; pp. 3122–3127.
19. Al Zaabi, M.R.; Alhendi, H.; Alkhoori, H.; Alkhoori, M.; Abu-Kheil, S.A.Y. Automatic site inspection system in construction sites (ICI-intelligent camera inspection). In Proceedings of the Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 21–24 February 2022; pp. 1–6.
20. Zhang, C.; Tian, Z.; Song, J.; Zheng, Y.; Xu, B. Construction worker hardhat-wearing detection based on an improved BiFPN. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8600–8607.
21. Nath, N.; Behzadan, A.H. Deep generative adversarial network to enhance image quality for fast object detection in construction sites. In Proceedings of the Winter Simulation Conference (WSC), Orlando, FL, USA, 14–18 December 2020; pp. 2447–2459.
22. Liu, Y.; Jiang, W. Detection of wearing safety helmet for workers based on YOLOv4. In Proceedings of the International Conference on Computer Engineering and Artificial Intelligence (ICCEAI), Shanghai, China, 27–29 August 2021; pp. 83–87.
23. Filatov, N.; Maltseva, N.; Bakhshiev, A. Development of hard hat wearing monitoring system using deep neural networks with high inference speed. In Proceedings of the International Russian Automation Conference (RusAutoCon), Sochi, Russia, 6–12 September 2020; pp. 459–463.
24. Pardhi, K.; Kokode, P.; Koche, S.; Thakur, P. Hard hat detection. *Int. J. Adv. Res. Comput. Commun. Eng.* **2022**, *11*, doi: 10.17148/IJARCCCE.2022.11444.
25. Kawade, Y.; Kasture, M.; Mallawat, Y.; Dubal, A. Construction safety equipment detection system. *Int. Res. J. Eng. Technol.* **2021**, *8*, 1581–1584.
26. Tyagi, R.; Thomas, K.T. Multiple safety equipment's detection at active construction sites using effective deep learning techniques. In Proceedings of the 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; pp. 1271–1276.
27. Zhou, M.; Fang, Z.; Zhao, B.; Li, P. Safety helmet wearing detection and recognition based on YOLOv4. In Proceedings of the 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 10–12 December 2021; pp. 798–802.
28. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. *arXiv* **2018**, arXiv:1801.04381.

29. Aryasa, K.; Rusydi, A. Design and build a sign language detection application with Tensorflow object detection and SSD Mobilenet V2. In Proceedings of the 5th International Conference on Cybernetics and Intelligent System (ICORIS), Pangkalpinang, Indonesia, 6–7 October 2023; pp. 1–5. <https://doi.org/10.1109/ICORIS60118.2023.10352247>.
30. T, R.; Mukherjee, A.; Sati, S. Face mask detection using SSD-Mobilenet-V2. In Proceedings of the 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 19–20 January 2023; pp. 366–372. <https://doi.org/10.1109/Confluence56041.2023.10048845>.
31. Alfarizi, S.; Salsabila, Z.; Bahrullah, S.; Susanti, H.; Cahyadi, W.A. Real-time cat detection system using MobileNet-SSD V2. In Proceedings of the 8th International Conference on Instrumentation, Control, and Automation (ICA), Jakarta, Indonesia, 9–11 August 2023; pp. 276–280. <https://doi.org/10.1109/ICA58538.2023.10273134>.
32. Hidayat, R.; Pramudika, R.M.V.; Budiman, C.; Barri, M.H.; Pangaribuan, P. River trash sorting system using SSD-MobileNet V2. In Proceedings of the Asia Pacific Conference on Wireless and Mobile (APWiMob), Bali, Indonesia, 10–12 October 2023; pp. 220–224. <https://doi.org/10.1109/APWiMob59963.2023.10365626>.
33. Kumar, S.; Kumar, R. Real-time detection of road-based objects using SSD MobileNet-v2 FPNlite with a new benchmark dataset. In Proceedings of the 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 27–28 February 2023; pp. 1–5. <https://doi.org/10.1109/iCoMET57998.2023.10099364>.
34. “Hard Hat Workers Dataset”. Distributed by Roboflow. Available online: <https://public.roboflow.com/object-detection/hard-hat-workers> (accessed on 22 August 2022).
35. Welcome to MaxusAI. Available online: <https://www.maxusai.com/> (accessed on 25 June 2022).
36. Kunz, B. The Story of CVflow. Available online: <https://www.ambarella.com/blog/the-story-of-cvflow/> (accessed on 25 June 2022).
37. “Cos Poszł Nie Tak,” Kwejk. Available online: <https://kwejk.pl/przegladaaj/3692403/5/cos-poszlo-nie-tak.html> (accessed on 26 July 2024).
38. “Man Tumbles on Icy Driveway and Nearly Cracks Head on Wall,” Daily Mail. Available online: <https://www.dailymail.co.uk/video/news/video-1861776/Video-Man-tumbles-icy-driveway-nearly-cracks-head-wall.html> (accessed on 26 July 2024).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.